The paradox of social interaction: Shared intentionality, we-reasoning and virtual bargaining

Nick Chater[1], Hossam Zeitoun[1] & Tigran Melkonyan[2]

1: *Warwick Business School, University of Warwick*

2: *Department of Economics, Finance and Legal Studies, University of Alabama, Tuscaloosa*

Abstract

Social interaction is both ubiquitous and central to understanding human behavior. Such interactions depend, we argue, on shared intentionality: the parties must form a common understanding of an ambiguous interaction (e.g., one person giving a present to another requires that both parties appreciate that a voluntary transfer of ownership is intended). Yet how can shared intentionality arise? Many well-known accounts of social cognition, including those involving "mind-reading," typically fall into circularity and/or regress. For example, *A*'s beliefs and behavior may depend on her prediction of *B*'s beliefs and behavior; but *B*'s beliefs and behavior depend in turn on her prediction of *A*'s beliefs and behavior. One possibility is to embrace circularity and take shared intentionality as imposing consistency conditions on beliefs and behavior; but typically, there are many possible solutions and no clear criteria for choosing between them. We argue that addressing these challenges requires some form of we-reasoning; but that this raises the puzzle of how the collective agent (the "we") arises from the individual agents. This puzzle can be solved by proposing that the will of the collective agent arises from a simulated process of bargaining: agents must infer what they *would* agree, were they able to communicate. This model explains how, and which, shared intentions are formed. We also propose that such "virtual bargaining" may be fundamental to understanding social interactions.

Key words. shared intentionality; coordination; social interaction; we-reasoning; mind-reading

We intuitively explain human behavior, including our own, by giving reasons that may justify that behavior (Davidson, 1963; Dennett, 1987). Our reason-based explanations of each other's actions are often multi-layered and complex. So, for example, when buying a used car in cash, we may explain why *A* carefully counts the money and holds it up to the light by her suspicion that *B* is dishonest. We may further explain that *B* feels affronted, because *A* is demonstrating a lack of trust. Going a step further, we can explain why *A* remarks, "I suppose I've got to go through this rigmarole!" to reduce *B*'s sense of being distrusted, by implying that *A* makes such checks as a matter of routine.

Many perspectives on social behavior elaborate on these intuitive, reason-based explanations of individual thought and behavior (e.g., Ajzen, 1991; Bandura, 1982; Fishbein, 1979; Mercier & Sperber, 2011; Perner, 1991). One direction for such elaboration, rooted in cognitive science, focuses on the rich structures of prior knowledge and experience that provide flexible "scripts" for common types of interaction (Minsky, 1974; Schank & Abelson, 1977). A second focuses on the complex role-playing and turn-taking conversational structure of many social interactions (e.g., Goffman, 1959; Haviland & Clark, 1974; Levinson & Torreira, 2015). A third focuses on the pragmatic inferences required to infer the meaning of linguistic and non-linguistic social cues (Clark, 1996; Grice, 1975; Sperber & Wilson, 1986). These and other projects go far beyond common-sense by extending and deepening, rather than replacing, reason-based explanations of social behavior.

Psychology as a discipline focuses primarily on the individual, and on the mental processes of which individual thoughts and behaviors are composed. Thus, psychological explanation typically, as in the example above, attributes reasons to individual people. From this point of view, organizations, social classes, and nations are not irreducibly ascribed "motives" or

"beliefs." Instead, any idea of reason or agency of a group should be explained purely in terms of the motives and beliefs of its members. It has generally been left to other social sciences, such as sociology, economics, management, and political science, to explain how group-level agency can arise from interactions between the individual human agents of which they are composed.

Recently, though, psychologists and philosophers have become increasingly interested in reason-based explanations that apply to more than one person, as embodied in notions of joint action, joint attention, and shared intentionality. We shall use "shared intentionality" broadly, to cover any cases in which people reach a commonly agreed interpretation, understanding, decision or plan.[1] Shared intentionality—including the ability to act, attend, and plan jointly—has been viewed as a key feature distinguishing humans from other primates (e.g., Call, 2009; Tomasello & Rakoczy, 2003). Moreover, shared intentionality has been seen as crucial to the development of culture (Searle, 1995; Tomasello et al., 2005); as playing a key role in cognitive and social development (Tomasello & Carpenter, 2007); supporting coordinated physical activities (Sebanz, Bekkering, & Knoblich, 2006); shaping and directing individuals' attentional and emotional resources (Boothby, Clark, & Bargh, 2014; Shteynberg & Apfelbaum, 2013; Shteynberg et al., 2014); motivating people to coordinate their efforts (Thomas, DeScioli, Haque, & Pinker, 2014); and providing a foundation for language (Clark, 1996; Clark & Brennan, 1991; Tomasello, 2008). Indeed, as we shall see below, social interactions of all kinds may often involve shared intentionality as a key element.

[1] Shared intentionality is the term preferred by Tomasello (e.g., 2005); in philosophy, the term collective intentionality is also widely used, though analyzed in a variety of importantly distinct ways (e.g., Bratman, 1993; Gilbert, 1989, 2009; Searle, 1990; Tuomela & Miller, 1988). In economics, related research has been developed under the heading of team reasoning (e.g., Bacharach, Gold, & Sugden, 2006). Yet further terms, including we-reasoning and we-thinking, are also widely used (Akerlof, 2016; Hakli, Miller, & Tuomela, 2010). Here, we use shared intentionality to cover all these lines of research. Schmid (2012) notes that related ideas about collective thought can be traced to figures as diverse as Juergen Habermas (1987), Wilfrid Sellars (1968), R. G. Collingwood (1947), and German phenomenologists in the early twentieth century. We will consider below how the specific account outlined, based on virtual bargaining, relates to some of these ideas.

In philosophy and economics, the question of how or whether shared intentionality can be explained in terms of the thoughts and behaviors of individual agents, has been viewed as a major theoretical challenge (e.g., Bacharach et al., 2006; Bratman, 1992, 1993; Gilbert, 2006; Searle, 1990). By contrast, in psychology, we suggest that the size of the gulf between reason-based explanation of single individuals and reason-based explanation of the *interactions* between individuals is often not fully appreciated.[2] Indeed, it is often assumed that the ability to "read" the minds of others may be sufficient, typically with the addition of so-called "higher-order" intentionality, recognizing that the others are also mind-readers. In this paper, we will argue that this is not the case.

The problem of "shared," or "joint," reasoning to reach some common understanding about each other's thoughts and behavior is, as we shall see, widespread. But it has clear boundaries: not all interactions between people have a fundamentally collaborative character aiming to establish a common understanding or plan. Consider, for example, situations in which one individual interacts with another purely as a physical "object," whether deliberately pushing someone out of the way, or inadvertently knocking them over. Equally, many aspects of social interaction are fundamentally one-directional: one party is attempting to interpret the behavior of another, ascribing them beliefs and motivations; or one person is attempting to control or manipulate the thoughts or behavior of another (e.g., so-called Machiavellian intelligence, widely discussed in primatology, e.g., Byrne & Whiten, 1990; de Waal, 2007). Indeed, much social behavior is fundamentally competitive, so that each party aims to understand and/or control the behavior of the other to their own advantage. In highly competitive interactions, the goal is typically not to reach a common understanding, but to outwit the other. Moreover, even

[2] There are notable exceptions, including Colman (2003); Colman, Pulford, & Rose (2008); De Freitas, Thomas, DeScioli, & Pinker (2019); and Shteynberg et al. (2020).

where a common understanding is crucial to successful interaction, this need not be obtained by joint reasoning between parties involved. A common understanding of the rules of the road, or the rules of tennis, or chess, or indeed linguistic conventions and moral norms, need not be created from scratch in interacting with another person, but may be presupposed as part of a shared culture.[3]

Yet collaborative, shared reasoning can often play a crucial role in even these cases. Although pushing another person does not involve collaborative, shared thinking, the evaluation of such actions (e.g., as impolite), and whether apology is required, or protest justified, may well do. If *A* pushes *B* out of the path of an oncoming vehicle that *B* had not noticed, there may be a joint expectation that A will be thanked rather than reprimanded. Similarly, in playing tennis, coordinating on when and where to play, or what is appropriate in a warm up involves shared intentionality; but in a competitive rally, each player will aim to disguise her intentions from, rather than share her intentions with, the other. Moreover, behavior is often guided by the interpretation of the rules (e.g., deciding whether a soccer shot is "over the bar" or "hits the post" where the goal is marked by two jumpers on a children's playground). Forming such tacit agreements seems paradigmatically to involve shared intentionality (e.g., Searle, 1995). Moreover, finding such shared interpretations will be particularly important in verbal communication in the light of the highly underspecified nature of linguistic meanings (Grice, 1975; Levinson, 2000; Sperber & Wilson, 1986; Turner & Horn, 2018) and the need to "anchor" those meanings in a specific context of people, objects, locations, and so on (Clark, 2021). But, once the rules are commonly established, playing the game itself is not a matter of shared

[3] Moreover, social coordination may often arise through the operation of shared cognitive and perceptual mechanisms, which align thought and behavior between people. Such mechanistic accounts, as in influential models of dialogue in psycholinguistics (e.g., Pickering & Garrod, 2004), lie outside the scope of reason-based explanation entirely.

intentionality. In chess, for example, the action of "launching an attack on the King" does not depend on mutual recognition by both players: indeed, the attack is more likely to succeed if the other fails to recognize it.

Shared intentionality arises where people need to share the interpretations of relevant actions; but it is puzzling how such a shared interpretation can arise. Specifically, attempting to reach such a shared interpretation runs into what we call the problem of *mutual prediction*. For person *A* to find a shared understanding of an action with *B*, person *A* needs to infer *B*'s understanding of that action. But *B*'s understanding will, in turn, be shaped by *B*'s attempt to infer *A*'s understanding. Thus, *A*'s thoughts and actions depend on *B*'s; and in turn *B*'s thoughts and actions depend on *A*'s, in an apparent loop.[4]

As we will see, this seemingly innocuous observation is surprisingly problematic, and indeed paradoxical, for reason-based accounts of social interaction involving the formation of shared intentions. We consider various attempts to address this challenge; and suggest a solution using the theory of virtual bargaining (Melkonyan, Zeitoun, & Chater, 2018, 2021; Misyak & Chater, 2014; Misyak, Melkonyan, Zeitoun, & Chater, 2014).

Outline of the paradox

Let us introduce the paradox with a simple example. *A* passes *B* a copy of *A*'s latest book. *B* has to determine the nature of this social interaction from ambiguous clues. Let us limit ourselves to two possibilities: that *A* may be *giving* (*G*) the book to *B*, in which case *B* should thank *A*

[4] We stress that there is nothing inherently paradoxical in recursion per se, which is, after all, familiar and central in computer science, cognitive science, and mathematics (Hofstadter, 1979). The problem is how such reasoning can resolve the problem of coordination between people who are operating independently. See, for example, Bacharach, Gold & Sugden's (2006) discussion of the so-called "hi-lo" game, where there are multiple equilibria and the need for some rationale by which people can coordinate on the same one.

profusely and keep the book; or *A* may be *showing* (*S*) the book to *B*, in which case *B* should examine it with interest, make suitably obliging remarks, and return it. A shared interpretation is crucial for harmonious social interaction. In the language of game theory (Schelling, 1960), social interactions require that *A* and *B* play a *coordination* game in order to resolve this ambiguity between *G* or *S*.[5]

At first glance, the challenge of making the correct inference may seem to lie entirely with *B*. That is, *B* has to engage in some process of mind-reading to establish *A*'s intention (*G* or *S*?). But note that *A* faces the complementary problem: predicting how *B* will interpret being handed the book.

For both parties, a successful communicative and social interchange requires that their interpretations, whether *G* or *S*, are the *same*. Indeed, considerable social confusion and embarrassment will result if *A* believes the book to be a generous gift, yet *B* hands the book back after a brief look and a few perfunctory remarks. Matters may be even worse if *A* intended *B* merely to inspect the book, but *B* gushes with thanks about *A*'s kindness and stubbornly fails to release it.

The two parties face a problem of mutual prediction. In rationally deciding on how or whether to pass the book to *B*, *A* must ask: will *B* interpret my action as *G* or *S*? But to decide this, *A* must ask what *B* will think *A* intends; and *B* will, of course, assume that *A* intends whatever *B* will infer from observing *P*. *A* and *B* seem to be reasoning in a circle: *A* is trying to "read" *B*'s mind; but *B* is trying to read *A*'s.

Thus, mutual prediction seems to lead inexorably to deadlock or infinite regress. We call this the *paradox of social interaction*. The paradox seems to arise widely. In many social

[5] The analysis here draws heavily on Clark (1996).

interactions, there will be the mutual challenge of inferring what the nature of the interaction is from ambiguous information: e.g., who is supposed to play which role and take which actions (e.g., Clark, 1996; Goffman, 1959). We stress that the paradox of social interaction is a challenge for theorists, rather than individuals. Specifically, the puzzle for theorists is how, given the apparent threat of deadlock or regress, people are able to navigate reciprocal interactions with other people so deftly.

Three points are worth stressing. First, any specific ambiguity in a social situation may, of course, be resolved, or at least reduced, by one or both parties. The person passing the book might present it ostentatiously, in order to clarify that it is a gift; or strongly imply that it is not (e.g., by commenting "you can have a look, if you like"). Nonetheless, most social situations will retain *some* measure of ambiguity. For example, even if it is made clear that the book is presented for inspection, there remains uncertainty about, for example, how rapidly it should be returned, how delicately it should be handled, or whether or not permission is required to scan its cover or opening pages with a smartphone. And explicit communication, however extensive, never resolves all ambiguities—this is parallel with the observation in legal theory that the law is inevitably "open-textured," and open to further interpretation (Hart, 1994). Indeed, natural language itself is notoriously riddled with ambiguities, from the scope of quantifiers, the extension of concepts, the reference of names, pronouns, and so on (Sennet, 2016). Thus, the challenge of coordinating on a shared understanding through coordination is not eliminated by the possibility of communication, but rather recurs in the interpretation of communication.

Second, disambiguation is itself a subtle and sensitive social action. *A*'s overzealous signaling that the book should be returned will imply a concern that *B* would otherwise be likely to keep it. *A*'s signaling may impugn *B*'s level of social competence by implying that *B* is liable

to make social gaffes; or impugn *B*'s motives by implying that *B* might be deliberately exploiting a potentially ambiguous situation to make off with the book. For such reasons, people will typically exercise caution in explicating what otherwise might reasonably be inferred to be the shared understanding by both parties.

Thirdly, note that agreeing on the nature of the social interaction seems to be required even where the interaction may be hostile. For example, even an overtly angry act is not thereby necessarily *hostile*. *B* might furiously tear up *A*'s book without concern for whether *A* is observing or even present. *B*'s act is hostile only when it is expressly aimed at conveying *A* and *B*'s shared recognition of *B*'s contempt for the book (and perhaps also for its author). Or suppose *B*, avenging some past slight, attempts to insult *A* by putting the book into the trash. *B*'s intention is that both *A* and *B* will agree that this action is an insult to the book's author, *A*. If *A* misinterprets *B*'s behavior as accidental, for example, it will not have the desired insulting effect.

What makes putting the book into the trash part of a hostile *interaction* is that *A* and *B* share the interpretation of the action as a deliberate insult. Thus, *B* is actively attempting to insult *A*; and *B*'s success requires that *A* recognizes this intention; that *B* recognizes that *A* has this recognition; and so on. In the view of the psychology of language, this understanding of the action as an insult is in *common ground* (Clark, 1996; we will discuss the puzzling notion of common ground further below). While common ground is a subtle notion, and one which we will use further below, we need only an informal understanding of the idea here: for some piece of information to be in common ground for *A* and *B*, we require that they both know it; know that they each other know it, etc.[6] From a mutual prediction perspective, it is hard to see how a

[6] Common ground might arise from publicly available information, shared culture, a common perceptual environment, collective attention (Shteynberg et al., 2020), or other sources (e.g., Clark, 1996). In economics and philosophy, the more usual term is common knowledge, and here we use these terms interchangeably, following, e.g., Thomas et al. (2014). The notion can be spelled out formally, either in set theory (e.g., Aumann, 1976) or

common ground interpretation of *B* putting *A*'s book in the trash can ever arise. *A* and *B* seem to be led down an endless regress in which each tries to infer how the other would interpret the action.

Social interactions, then, frequently involve generating shared intentions and shared understanding from potentially ambiguous cues. This, in turn, seems to generate an apparently paradoxical interdependence between *A* and *B*: the thoughts and actions of each seem to depend on the thoughts and actions of the other. How, then, might we attempt to escape the paradox, and deal with the mutual interdependence of *A* and *B*'s actions?

One approach is to embrace circularity: to note simply that there are two "readings" of the action which are consistent and, if adopted, will lead to harmonious interaction. On one reading, *A* and *B* both interpret the physical movements of conveying the book from one person to the other as an instance of *S*; both believe that the other adopts reading *S*, that the other believes that the other believes that they adopt reading *S*, and so on. On the other reading, both interpret the same physical movements as a case of *G*, with the parallel hierarchy of beliefs, and beliefs about beliefs. But the paradox remains: how do the parties know *which* interpretation to agree on?

It is tempting to suppose that solving coordination problems is, in practice, often made easier by a clue or hint. Indeed, *A* might wave the book towards *B* very casually, rather than slowly and formally, perhaps seeking to imply, "this is no big deal—just have a quick look," and this might seem to imply that *A* does not intend to give the book as a gift. But notice that any hint will work only if both parties *agree* on the interpretation of the hint (e.g., that it favors *S* rather

epistemic logic (e.g., Harman, 1977; Schiffer, 1972). There is a debate concerning whether Lewis' (1969) important early formulation is a distinct approach (Cubitt & Sugden, 2003; Vanderschraaf, 1998). Our discussion here does not depend on these fine details.

than *G*). But then both parties face a new coordination problem: coordinating on the interpretation of the "hint," and the paradox is just as problematic as before.

Why has the paradox been largely unnoticed in psychology?

We have suggested that the paradox of social interaction is fundamental. But discussion of this, and related, issues is sparse in psychology. There are notable exceptions, including analysis of these issues in the context of language and communication (e.g., Clark, 1996), and in social (Shteynberg et al., 2020) and developmental (Bohn & Koymen, 2018) psychology. Moreover, there is work showing the power of common ground in experimental games (Colman, 2003; Colman & Gold, 2017; De Freitas et al., 2019; Thomas et al., 2014); and proposals concerning the need for evolved cognitive mechanisms to think "as a group" (Tooby & Cosmides, 2010). But given that the centrality of social interaction is so central to human life, it is natural to ask why the paradox has not been a major focus of psychological discussion.

One reason is, we suspect, that humans are so good at resolving such coordination problems in practice that their very existence is not noticed: the "right" interpretation of an action (e.g., a gift, an insult) seems so obvious that it may seem to require no further explanation. Another reason is that a great deal of experimental research, even in social psychology, does not involve *interaction*: indeed, for good reasons of experimental control, paradigms involving the interplay of freely interacting agents are often avoided where possible.[7]

In short, social interaction is often not the immediate focus of investigation in social psychology and related fields. For example, considerable interest has focused on what we might

[7] Some authors have argued that experimentally studying social and linguistic *interaction* is of crucial importance. They have not necessarily endorsed the reason-based style of explanation advocated here (Pickering & Garrod, 2004, 2021; Schilbach et al., 2013).

term *social perception* (e.g., Heider, 1958; Kelley, 1967; Macrae & Bodenhausen, 2000) or *mind-reading* (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Nichols & Stich, 2003; Singer, 2006). Here people infer another person's mental state from their observed behavior, but where the person observed is not simultaneously making an inference about the observer. Similarly, there is considerable interest in how a person's thoughts and behavior are affected by people and groups around them. *Social influence* (e.g., Asch, 1956; Cialdini & Goldstein, 2004; Sherif, 1936) too, need not be reciprocal (although it can be: Mahmoodi, Bahrami, & Mehring, 2018). A person's thoughts and behaviors will be shaped by those around them, irrespective of whether the converse is true. Another area of intensive research interest is the *social transmission* of thoughts and behaviors through populations and across generations (e.g., Boyd, Richerson, & Henrich, 2011; Christakis & Fowler, 2009; Heath, Bell, & Sternberg, 2001). The spread of linguistic expressions, gestures, motor skills, levels of aggression, food and alcohol consumption, norms of politeness, emotions, rumors, and beliefs is typically considered as involving one-way transfer from one agent to another. In each of these important domains, the relationship between one agent and the next can be modeled as one-directional, without needing to invoke mutual prediction. Table 1 provides an illustrative sample of well-known experimental studies in social, developmental, and comparative psychology, under this classification.

Insert Table 1 about here

It is perhaps natural to hope that a theory of social interaction might be created by adding a one-directional influence from *A* to *B* to a one-directional influence from *B* to *A*. But the paradox of social interaction illustrates how mutual influence can introduce wholly new

phenomena: people must simultaneously and independently infer relevant beliefs, intentions, and actions for both parties.[8] To the degree that researchers have focused on one-way interactions, and anticipated a natural generalization to the two-way case, the paradox of social interaction has been largely unnoticed.

The paradox and some current theoretical accounts

Many influential reason-based theories of social development, cognition, and behavior appear either (i) to run into the paradox; or (ii) to leave it unresolved. In the first category are process models that, implicitly or otherwise, propose that person *A* attempts to mind-read person *B*, taking account of the fact that *B* is attempting to mind-read person *A*, etc. In the second category are models that embrace circularity by defining "equilibrium conditions" under which *A*'s and *B*'s beliefs and/or actions can be consistent. But, as we noted above, this side-steps the crucial question—how do *A* and *B* coordinate on *which* of multiple equilibria to choose (do they both see the handing over of the book as giving or showing?).

The first problem (circularity) applies to prediction-based models, which are widespread in the cognitive and brain sciences (Clark, 2013), and have recently been applied to social behavior (e.g., Tamir & Thornton, 2018). According to this viewpoint, the brain is a "prediction machine" with a fundamental drive to learn by reducing its prediction error. But in the context of mutually interdependent social interaction, this approach treats social interaction as a problem of mutual prediction, and our paradox hits with full force. If *A* and *B* are two "prediction machines" each attempting to predict the other, they are seemingly unable to avoid an infinite regress.

[8] We leave open the possibility that social perception, influence, and transmission may sometimes involve some degree of mutual interaction—indeed, this will be the norm in communicative contexts such as our 'book donating' example above. Our point is simply that they need not do so.

The same issue recurs, independent of the mechanism used to make such predictions. For example, consider the simulation-theory of mind (e.g., Goldman, 2006; Gordon, 1986), according to which people attempt to mind-read by simulating the thinking of others using their own knowledge and reasoning. The essential claim of the simulation-theory is that, rather than requiring a "theory" or "model" of the mind of the other, the other's mind can be predicted by using one's own mind as an analog of the other's mind. Consider, for example, a criminal wondering how to mislead a detective by leaving false 'clues.' According to the simulation theory, to assess what a detective will infer from seeing a clue, the criminal asks herself what she herself would infer from seeing that same clue. This might involve modifying her own beliefs appropriately to align with those of the detective. Thus, the criminal might ask what, hypothetically, she would infer from a clue, were she not already aware of her own guilt; or what she would infer if, like the detective, she were convinced that no crime could be committed by a judge or a member of the clergy. Whatever the attractions of this account, it has limited purchase on the paradox of social interaction. In such situations, each person is attempting to simulate the other. Thus, *A* needs to simulate *B* simulating *A* simulating *B*, and it is not clear how infinite regress can be avoided. The simulation theory just embodies the paradox, rather than helping to resolve it.

The picture is no different for the influential theory-theory of mind (Gopnik & Wellman, 1992; Wellman, 1992), according to which people infer the mental states of others to best explain their behavior. That is, an observer builds a "model" of the thoughts of the other. This approach works well for social perception, where one person observes, but does not interact with, another person. Indeed, this viewpoint has been productively modelled by extending Bayesian methods that have been so successful in capturing the perception of the physical world

(e.g., Knill & Richards, 1996; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Yuille & Kersten, 2006) and social behavior (Baker, 2012; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009).

In social interaction, though, this recursive approach runs into a version of our now-familiar paradox. Consider *A* and *B*'s interaction above. According to the theory-theory viewpoint, to understand *B*, *A* must have a model of *B*'s reasoning. But conversely, to understand *A*, *B* must have a model of *A*'s reasoning. This seems to imply that *A*'s model of *B*'s reasoning must contain a nested model of *A*'s reasoning, which presumably in turn will include a further model of *B*'s reasoning—we have our usual problem of infinite regress. Most psychological accounts of this type implicitly or explicitly assume that such recursion is truncated (e.g., Tomasello, 2008). That is, people model one another's beliefs ("first-order" beliefs); perhaps also model that others model their own beliefs ("second-order" beliefs); and perhaps occasionally extend to the third and fourth order (Liddle & Nettle, 2006).

A variant of this approach is also embodied in formal theories in behavioral game theory (cognitive hierarchy theory (Camerer, Ho, & Chong, 2004) and level-k reasoning (Costa-Gomes & Crawford, 2006; Nagel, 1995; Stahl & Wilson, 1994)). This truncated reasoning approach, while helpful in analyzing certain kinds of competitive interactions, has difficulties in dealing with coordination. Consider, for concreteness, the case where *A* and *B* are both second-order reasoners. *A* then tries to do whatever she thinks *B* thinks he (*A*) will do.[9] But knowing what *B* thinks *A* will do is at least as difficult as deciding what *A* will do herself. And *B* faces a similar challenge, trying to establish what *A* thinks he will do, when he has yet to decide this himself. Of

[9] The problem is, in more technical terms, that developing a theory of Level-0 reasoning (which is presumed to be the starting point for higher level reasoning) is in these cases just as difficult as solving the original problem of deciding what the players will do. In practice, Level-0 reasoning typically, though not always, corresponds to a uniform probability distribution over the set of all strategies available to a player.

course, agents might use some heuristic to infer these difficult second-order beliefs. But, if so, then (a) the entire explanatory burden falls on the heuristics, not the theory of mind-reading; (b) if such heuristics existed, both parties should surely apply them directly to determine their own behavior, and might avoid becoming lost in unnecessary recursive reasoning; and (c) the question of how the heuristics emerge to lead to successful coordination is entirely unaddressed. In sum, theories that assume that people's recursive mind-reading is limited to a finite number of levels cannot easily explain how people coordinate successfully in social interactions.[10]

The second category of models, as we have noted, embraces circularity, and seeks to define a notion of equilibrium, in which the actions and/or beliefs of the parties are consistent (although then facing the puzzle of how people are able to coordinate by choosing the same equilibrium). This approach is standard in game theory in the form of the Nash equilibrium and its many variants. It is also embodied in some sophisticated Bayesian computational models of social reasoning in psychology. For example, consider the social interaction of "optimal" teaching in which an agent aims to present the information that will be most useful to a learner. Shafto, Goodman and Griffiths (2014) define Bayesian equations in which teachers (approximately) optimize their choice of an action in the light of their model of the listener; and listeners optimize their interpretations of the information from the teacher, based on their model of the teacher; and use iterations to enhance their consistency. This approach has been extended to deal with pragmatic reasoning, with the creation of rational speech act theory (Frank & Goodman, 2012; Goodman & Frank, 2016). This work is interestingly related to the notion of

[10] Where levels of mind-reading are evaluated directly, recursive depth is assumed to be very limited. For example, behavioral games, such as the "beauty contest" game (Nagel, 1995), are used to measure recursive depth in competitive interactions. They typically suggest that people may best-respond to the other (i.e., first-order reasoning), or best-respond to a first-order reasoner, but that higher-order reasoning is rare. As a reviewer has pointed out, obtaining predictions from level-k reasoning is complicated further by the fact that depth of reasoning differs between people, and, indeed, is not consistent across games (Georganas, Healy, & Weber, 2015).

"rationalizability" in game theory (Bernheim, 1984; Pearce, 1984), in which players' choices can be rationalized as a best-response in the light of some beliefs about the other players (though these beliefs need not be correct). As Shafto, Goodman and Griffiths (2014) note, their approach, like rationalizability, typically generates many possible solutions. Therefore, some additional mechanism is required to select a specific prediction.

One particularly interesting approach along these lines, which can be viewed as formalizing the theory-theory approach, is "action interpretation as inverse planning" (Baker et al., 2009; Wang et al., 2021). The aim is to provide a reason-based account of social perception by applying a Bayesian analysis-by-synthesis approach, widely used in modelling perception (Yuille & Kersten, 2006) and language interpretation (Bever & Poeppel, 2010), to the problem of understanding the intentions behind observed actions. The approach begins with prior assumptions about the rational planning of actions in the light of goals and beliefs; and uses Bayesian inference to invert this model and hypothesize on the goals, beliefs, and plans from the observed action. This work provides an elegant account of how it is possible to "read" complex beliefs and preferences from behavior. For example, if we see an agent circumspectly skirt an area of ground while heading towards a goal, we may tentatively infer that the agent believes there is some hidden obstacle or danger that would be encountered on the direct route—otherwise their behavior would not correspond to an optimal plan.

Like other approaches to social perception, this strategy would lead to deadlock if directly applied to social interactions: *A*'s optimal plan depends on her inference about *B*'s optimal plan, which in turn depends on *B*'s inference about *A*'s optimal plan. The paradox arises because, in social *interaction*, plans are entangled.

Insert Table 2 about here

We-reasoning: social interactions and joint actions

There is a natural and intuitive way to escape the problem of mutual prediction. According to this approach, we should not imagine each person asking: *what do you think that I think that you think…* without limit. Instead, each participant in an interaction should ask: what do *we*, considering ourselves composing a single supra-agent (a team, a group), think we should do? Crucially, this perspective involves assigning beliefs, desires, and reasoning processes, not merely to individual members of a group or team; but to the group or team considered as a single agent. Indeed, extending talk of mental states not just to individuals but to groups is familiar in everyday discourse. We speak of a jury believing a criminal to be guilty; a parliament deciding to pass a law; or Real Madrid aiming to play entertaining rather than defensive soccer. In such contexts, we are saying more than that each individual in the relevant group has a particular view (and indeed, individual jurors, for example, might privately dissent). We instead view the group as a single collective agent (Gilbert, 1987).[11]

Using this viewpoint, then, one way to address the problem of mutual prediction is (i) for the two or more interacting individuals to consider themselves as a single collective agent of which they are a member; (ii) for each individual to ask how this collective agent should reason

[11] A reviewer has pointed out that the sociologist Max Weber influentially raised skepticism about ascribing beliefs, intentions, or choices to groups or organizations, unless these can be recast in terms of the beliefs, intentions, or choices of the individuals of which these organizations are composed—on the grounds that only individual people have a subjective understanding (*Verstehen*) of their actions. How ascriptions of joint mental states can be ascribed to a group based on the mental states of its members has been an important focus in psychology, philosophy, and economics, and one to which we return below.

and what it might decide; and (iii) for each agent to choose their individual courses of action according to the imagined recommendations of the collective agent.

This type of approach is often referred to as *we-reasoning*, and has been widely discussed and developed in related but distinct ways, across philosophy, economics, and psychology (e.g., Bacharach et al., 2006; Gallotti & Frith, 2013; Gilbert, 2006; Hakli et al., 2010; Sugden, 2003; Tenenbaum et al., 2011). Indeed, as noted above, we-reasoning, under the label of "shared intentionality," has been proposed as a distinctive feature of human cognition (e.g., Tomasello & Rakoczy, 2003). Of particular interest, in terms of underlying cognitive mechanisms, is the concept of shared or joint attention, in which it is common ground that people are attending to the same object or event. An extensive program of research in social psychology has found that people preferentially devote cognitive and emotional resources to items they believe to be "co-attended" with others (e.g., Boothby et al., 2014; Shteynberg & Apfelbaum, 2013; Shteynberg et al., 2014). Shteynberg et al. (2020) develop this viewpoint further, with notions of collective attention and collective learning, which seem to be crucial to coordinating group behavior. These ideas help clarify the underpinnings of we-reasoning: for a group of people to share thoughts and intentions, they need minimally to be collectively attending to the same information and collectively attempting to solve the same problem.

The notion of we-reasoning does, though, raise a substantial theoretical challenge: How can an apparently mysterious collective agent arise from the thoughts and actions of individual agents?

One clue concerning how to proceed is to consider the status of the conclusion that is sought: for example, that given relevant background knowledge, situational factors, and the nature of the action, *we* should agree that this is an act of, say, giving (*G*) rather than showing

(*S*). But presumably *we* can only derive conclusions based on premises that *we* endorse: i.e., from information that is in common ground (Clark, 1996).[12]

So, for example, suppose that it is in common ground that *A* just said to *B*: "I'll be really interested in your thoughts about my new book," and handed *B* the book after signing it with a fountain pen. In the light of this information, it seems overwhelmingly likely that *A* is giving the book to *B*. Indeed, this would surely be the presumption of a disinterested third party observing the interaction. If this inference is itself in common ground, then the interpretation *G* will also be in common ground. Thus, *B* can safely thank *A*, and hold on to the book, without fear of social disaster.

The reason-based explanation that spells this out more fully turns out to be surprisingly subtle. To interpret the intention behind *A*'s action, *B* must ask: in the light of the knowledge of this situation that is in common ground for both of us, what is the natural interpretation of this action? That is, what would a disinterested third party, sharing this common ground, conclude about *A*'s intention (*G* or *S*?). Similarly, in generating the action, *A* should follow the same reasoning to assign an intention to her own action. Thus, in choosing her action with the aim that the social interaction will run smoothly, *A* must ask: what action can I take such that we would both agree that the intention is *G* (or *S*, depending on whether *A* wants to give, or merely show, the book to *B*).

One way to put this point is to see social interactions as a type of shared mental activity (even if the action, such as giving or showing, is conducted by only one partner).[13] There may, of course, be additional layers of social complexity, which will be outside the scope of we-

[12] Indeed, if one party uses private knowledge not available to the other person to choose their interpretation, this will likely disrupt coordination, because the other cannot "follow" the same line of reasoning.

[13] The viewpoint requires a 'rich' notion of joint action, which we endorse; but some theorists propose more minimalist accounts (e.g., Vesper, Butterfill, Knoblich, & Sebanz, 2010).

reasoning. For example, in handing over the book for inspection, *A* may hope to impress *B*, but probably does not want *B* to realize this. Similarly, *B* may infer that *A* is self-important, but assume that this is definitely not in common ground. Or *A* may suddenly fear that *B* will see her actions as boastful. The shared mental activity, though, is what makes this a social *interaction*, rather than simply one person attempting to manipulate or understand another.

The behavior arising from such shared mental activity can then be interpreted via inverse planning that we described above (Baker et al., 2009), but now at the level of the collective agent. Thus, the intention of the action (say, *G*) is derived by inferring what we (i.e., the collective agent with the common ground of *A* and *B*) would have to jointly intend, were we to decide to choose that *A* perform this action. Given that *A* has performed this action, we can then infer from common ground that we must have had that intention. In particular, if *B* can infer that this intention is in common ground (in the light of the observed action—e.g., handing over the book in a particular manner), then *B* can infer that *A* has that intention *G*. Anything that is in common ground must be known to both *A* and *B*; and hence, can be inferred by *B*.

From this "joint planning based on common ground" standpoint,[14] we can explain *why* certain features of the action would suggest one interpretation or the other. For example, opening the book and writing something in it makes sense according to the plan, "*A* signs and gives a

[14] This viewpoint is closely linked with important recent computational work in a Bayesian framework on a somewhat different, and in some ways more challenging problem (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016; Wang et al., 2021). In this work, the agents do not have the "payoffs" of possible combinations of actions in common ground, but must learn them from experience; and they must also learn the "utilities" of others, and whether the others are collaborating for the benefit of the team, or are maximizing their own rewards (Kleiman-Weiner et al., 2016). More recently, Wang et al. (2021) propose an interesting algorithm, Bayesian delegation, in which each agent attempts to infer the tasks of the entire set of agents using Bayesian inference from observing their actions. The payoffs to each agent depend on the interaction of different agents on a task (e.g., collaborating on some subtasks, distributing other subtasks between themselves). Here, the focus is on learning about, and hence coordinating appropriately, with other agents, by observing their actions over long periods of interaction. By contrast, the approach that we develop below, virtual bargaining, focuses on the complementary problem of how agents can instantaneously find a common plan, by reaching an agreement through a process of hypothetical bargaining.

book to *B."* But it is utterly unnecessary in the light of the plan, "*A* shows *B* a book for a moment." By contrast, pointing out a curious detail of the cover while handing over the book is a natural part of a plan in which *A* allows *B* to inspect a small and difficult-to-see quirk of the cover design. But it is unnecessary in the action of making a gift. Although explicating this reasoning is conceptually subtle, the degree to which it is reliable can in practice be assessed straightforwardly. Any disinterested third party with an understanding of the relevant common ground can decide how plausibly the conclusion follows.

The focus on the importance of common ground provides a basis for a reason-based understanding of some of the subtleties of everyday social interactions. Suppose, for example, that *C* whispers to *B* that *A* is giving the book away to everyone and that *B* should be ready to receive it. In this case, *B* now knows *A*'s intention to give the book to *B,* and *A* of course also knows her own intention to do so. But the intention is not in common ground: for example, *B* does not know that *A* knows that *B* knows it. And this is crucial. Suppose that *A* were to casually pass a copy of the book to *B*. *B* cannot then say, "Oh, such a generous present!" —indeed, this will be viewed by *A* as presumptuous, and by *B* as a terrible faux pas. Rather, *B* must wait until *A* signals that the gesture is a gift that *will* be in common ground, so that the information to underpin the conclusion that "this is a gift" can be inferred by we-reasoning. Furthermore, suppose that *A* overheard *C*'s whisper. Nonetheless, *A* will feel consternation at *B*'s claiming the book "all too readily"—because although *A* knows that *B* is expecting the gift, *B* behaves as if this expectation is in common ground (i.e., derived from we-reasoning), where it is not.

The invocation of common ground, rather than layers of mind-reading the other who is mind-reading us, breaks the circularity of the paradox of social interaction. But it raises puzzles of its own. In particular, from an individualistic perspective, the challenges of deciding

definitively whether some of the knowledge, *K*, is in common ground might seem to involve a particularly complex mind-reading challenge: *A* might seem to need to establish that both she and *B* know *K*, that they both know each other know *K*, and so on. If common ground is viewed as depending on an infinite hierarchy of beliefs about beliefs about beliefs, then establishing common ground itself sinks into paradox.

How can we possibly establish an infinite sequence of nested epistemic claims? In line with other authors (e.g., Clark, 1996; Shteynberg et al., 2020), we stress that common ground cannot be rooted in inferences from individual knowledge. For example, Clark (1996) notes that we may sometimes directly infer information common between us (e.g., an announcement over a public address system; a salient object that is in plain view for us both; or "well-known" facts, e.g., that we learn at school; norms routinely used by everyone in our community; or facts that anyone in our company must have learned in their induction, etc.). Thus, common ground can often be inferred directly, and the infinite sequence of nested epistemic claims can then be inferred as a consequence. Thus, *A* may assume that she and *B* have common ground that Paris is the capital of France, from the fact that capital cities are routinely learned at school; and then can infer that *B* knows this (because *A* and *B* were both present in the class when the teacher informed them that Paris is the capital of France, and they observed each other paying attention), that *B* knows that *A* knows this; that *B* knows that *A* knows that *B* knows it, and so on.[15]

Virtual bargaining

We noted earlier that the social sciences have a long tradition of explaining the behavior of groups in terms of interactions between the individuals of which they are composed. Indeed, this

[15] We thank an anonymous reviewer for raising this point.

viewpoint is particularly strong in psychology, where social behavior is explained through the lens of individual minds. From this point of view, a crucial theoretical challenge is to establish the preferences or goals of any postulated supra-personal entity (the "we" in we-reasoning), perhaps from the preferences and goals of the individuals of which it is formed.[16]

One approach is to assume that the "supra-agent" is benevolent towards its members, perhaps aiming to maximize their summed utilities (e.g., Bacharach et al., 2006; Colman & Gold, 2017). This approach faces the problem of interpersonal utility comparisons (Diamond, 1967; Elster & Roemer, 1993; Weymark, 2016) as well as the difficulty of taking into account the possibility that individuals never completely abandon their own objectives for those of the group.

A more fundamental problem arises from social interactions in which power is very unbalanced. Suppose, for example, *X* is a gangster and *Y* a downtrodden and fearful side-kick. If *X* passes a valuable stolen watch to *Y*, *Y* will interpret this as "showing" and return it right away. But if *Y* hands a stolen watch to *X*, *X* will keep it. Of course, if *X* is very rich and *Y* is very poor, summed utilities might be maximized if the watch is owned by *Y* not *X*; and a 'benevolent' team might thus choose to assign the item to *Y* (Bacharach et al., 2006). But this will be irrelevant for understanding the social interaction between the gangster and side-kick: this particular 'team' will agree that the watch will go to the less-needy but all-powerful gangster. Power relations, although of huge importance in understanding social interactions (Guinote & Vescio, 2010), are orthogonal to questions of unconditional aggregate utility maximization.

A further complication arises when, as is often the case, *A* and *B* must both coordinate their behavior but have partially conflicting interests. Suppose *A* and *B* are at a beach-side

[16] Some theorists suggest that such a reduction may not be possible (e.g., Searle, 1990)—that we-thinking needs to be taken as primitive, and not reduced to the mental states of individuals. We shall see shortly that virtual bargaining can be viewed as a particular way of reconciling a distinct "plural subject" (Gilbert, 1989) as being both distinct from, but compatible with, the minds of individual participants (Bratman, 1992, 1993).

restaurant with magnificent sea views. A square table at which they must be seated has one chair facing the sea with an excellent view (Good), but the opposite chair has no view at all (Bad). The other two chairs face side-on, both having a partial sea view (Medium1 and Medium2). Let us assume, plausibly, that *A* and *B* must sit opposite to each other (the table is quite small). So, either one person has Good and the other Bad; or they both get Medium outcomes.

Insert Figures 1 and 2 about here

Note, first, that models based on the Theory-Theory or Simulation-Theory of Mind, and prediction-based models, will be at a loss to deal with this scenario: the paradox is in full force. Either theory of mind-reading will propose that *A* tries to model *B*'s reasoning (presumably to take the opposite chair, to avoid social disharmony); but then *B* will be attempting to model *A*'s reasoning—and we have gone in a circle. And according to prediction-based models, both people are engaged in a deadlock of mutual prediction.

How far does we-reasoning take us? Let us start by considering what can be inferred from common ground. There are three possible joint outcomes, with payoffs [U_A=H; U_B=L], [U_A=L; U_B=H] and [U_A=M; U_B=M]. So, *A* and *B* are facing a coordination problem: ideally, they would identify the same option and smoothly instantiate it. But can they avoid choosing incompatible options leading to social confusion and embarrassment?

We suggest that the following line of argument presents a promising theoretical direction. Notice that deciding which of the outcomes is most appropriate requires a mechanism for trading

off *A*'s and *B*'s conflicting interests. If *A* and *B* were to communicate, then they might be able to decide on a trade-off by negotiation.[17] But in many cases it may be sufficiently obvious to *A* and *B* what conclusion this negotiation would reach that communication is unnecessary. Thus, we have a candidate mechanism to guide social interaction in the absence of communication—both parties need merely implement what they would have negotiated. We have elsewhere called this mechanism *virtual bargaining* (e.g., Melkonyan et al., 2018, 2021; Misyak & Chater, 2014; Misyak et al., 2014).

What, then, is the basis for the judgments concerning the "obvious" agreement? Note that, if such an agreement is to follow from reason-based explanation, any conclusion concerning the outcome of virtual bargaining must be in common ground: both players must be able to infer that they both know it, know that the other knows it, and so on. Imagine, in our sea view example, that *A* and *B* are rival business people, with no fellow-feeling. Nonetheless, person *A* has the "trump card" of living in the middle of a continent, and not having seen the sea for many years; whereas person *B* is a local who sees the sea every day. Both would prefer the sea view; and each has researched the other's background—so both know about *A*'s trump card. Nonetheless, if this is not in common ground (i.e., they do not know what background research the other has done), then it would be socially highly inappropriate for *A* to make a dash for the chair with the sea view, and *B* would react with consternation.

Reflection on examples of this kind suggest that virtual bargaining must derive the conclusion that "if we could communicate, we would agree to *this*" from common ground. The

[17] A reviewer has pointed out that deadlocks can still occur, even when communication is available, such as when two people, not sure who is next in the checkout line, end up politely trading "You go first!" "No, you go!" "No, really, you go ahead" "No, I'm sure you were here before me" sometimes to the point of mutual irritation.

decisive inferential step is moving from this common-ground hypothetical to each carrying out the inferred action.[18]

As in the examples above, what counts as the "natural" bargain in the light of common ground can, in general, be assessed by third parties, given that common-ground information. Indeed, when *A* and *B* are assessing what bargain they might hypothetically reach, they are attempting to adopt such a dispassionate stance. If they are able to do this successfully and a unique natural bargain emerges, then a harmonious interaction is likely to result. In practice, a full range of cognitive shortcomings (e.g., self-serving biases; projection of one's own interests to the other; the curse of knowledge) have the potential to derail this process and may lead to a dissonant social interaction. But, we suggest, social interactions can be understood as striving towards the objective of implementing a virtual negotiation based on common ground. Indeed, this framework is especially helpful in understanding attempts to forestall or repair dissonant interactions, for example, by *A* saying "wow, I've not seen the sea for 20 years" as *A* and *B* walk towards their table or stand at the table struggling with the choice of seats.

Note, too, that the virtual bargaining viewpoint naturally captures the impact of power asymmetries. If gangster *X* and side-kick *Y* approach the restaurant table, *X* will head directly for the best seat, and *Y* for the worst. Both can infer that this would be the outcome were they to bargain (i.e., *X* has overwhelming bargaining power); hence they simply implement the result of this hypothetical bargain without needing to communicate.

[18] A fruitful line of future research is to consider principles of counterfactual or hypothetical thinking observed in other domains, whether hypothetical reasoning about the physical world, or about individual behavior (Byrne, 2005; Roese, 1997). For example, we might expect that bargains that involve easily imagined changes from the status quo will be preferred (Kahneman & Miller, 1986); that people will consider only a very small number of possible bargains (Byrne, 2005; Johnson-Laird, 1983); and that they will be particularly likely to alight on salient or focal bargaining solutions (Schelling, 1960). Perhaps especially intriguing is the finding that social power may shape how readily people imagine that they might act differently (Scholl & Sassenberg, 2014). It is interesting to wonder whether this effect will carry over to tasks involving joint planning through virtual bargaining.

How far can virtual bargaining be modelled formally? In specific contexts, where common ground, including payoffs, is well-defined (e.g., in laboratory experimental games or economic transactions), this is possible. Indeed, we have developed a game-theoretic model, with equilibrium conditions ('feasible agreements') that possible bargains must fulfil (e.g., Melkonyan et al., 2018, 2021; Misyak & Chater, 2014; Misyak et al., 2014). To choose between feasible agreements and alight where possible on a single virtual bargain, the model needs to draw on a theory of bargaining (after all, this is what virtual bargainers are attempting mentally to simulate). In economics, there are various theories of bargaining where the "best" bargain depends only on well-defined "utilities" of the players. Melkonyan et al. (2018, 2021) use the most well-known theory, Nash bargaining[19] (Nash, 1953), which selects the option that maximizes the product of utility gains for the players for the bargain (over and above the utilities that would be obtained if bargaining failed). But other theories of bargaining could be used instead.[20]

But as for other aspects of informal reasoning and argumentation, which operate in an open-ended and only partially understood world, complete and precise formal analysis of virtual bargaining will often not be possible. Indeed, as in common-sense reasoning more generally (e.g., Fodor, 1983; Harman, 1986; Oaksford & Chater, 2007), the range of factors that can shape the natural bargain is limitless. Nonetheless, our intuitions about what bargain would be reached in a hypothetical negotiation are often quite clear cut. Consider the following variants of the chair-choosing scenario:

[19] Note that Nash bargaining is not connected to the Nash equilibrium mentioned above.

[20] Other theories of bargaining could also be used, e.g., Kalai-Smorodinsky bargaining, which focuses on a fairness criterion embodied in equalizing the ratio of utility gains (Kalai & Smorodinsky, 1975).

i. Person *A* is the CEO, and *B* a new employee. The natural agreement will probably be that *A* takes the better seat, and *B* the worse; and this may be implemented without comment. Interestingly, the CEO can "reopen" the negotiation to her potential disadvantage, e.g., by edging towards, or just selecting, one of the "Medium" chairs. For the new employee to do so would be a social gaffe).

ii. *A* and *B* are good friends; it is in common ground that neither will be happy if the other is disadvantaged. They may naturally both select "partial view" seats.

iii. *A* and *B* hardly know each other but are keen to make a good impression (i.e., to preserve social harmony). Thus, both know that it will be a faux pas to try to "capture" the best seat (i.e., they do not have sufficient common ground to reach any particular asymmetrical outcome); and hence they both align on the "partial view" seating arrangement (here, unlike in ii., there is no concern about the other's well-being). Now it is possible that either *A* or *B* may, like the CEO in i., attempt to renegotiate to their own disadvantage, e.g., by gesturing the other towards the seat with the sea view. If accepted, this may be interpreted as a friendly act. On the other hand, it may be viewed as overly ingratiating and insincere, perhaps as attempting to imply concern for the other's well-being, which seems to have no basis.

The aim of our arguments is to illustrate the *structure* of the social reasoning in which we routinely engage—the types of considerations that matter, the importance of what is in common ground, and the we-reasoning and virtual bargaining through which people strive to create implicit agreements underscoring harmonious social interactions. Once this structure is clear, it is possible to define precise formal models and make experimental predictions for specific

situations, based on assumptions about the social participants' common ground (of, e.g., their action alternatives and their payoffs) (e.g., Melkonyan et al., 2018, 2021).

Implications

We have argued that social interaction typically involves a mutual adjustment of thought and behavior, and that a reason-based explanation of this process cannot operate by a process of mutual prediction, on pain of circularity. Instead, coordination is achieved, where it can be achieved at all, by a process of simulated bargaining between the people involved in an interaction, where the outcome of that simulation is based only on their common ground (and is not any knowledge that is provided to one person). Here, we consider implications and applications of the paradox and its solution using virtual bargaining.

The paradox of social interaction and theories of shared intentionality. Psychological theorizing has drawn extensively on philosophical analyses of shared intentionality and related concepts. How does the present discussion, both of the paradox of social interaction and of the proposed solution via virtual bargaining, relate to such analyses? Most fundamentally, the focus here is on how people are able to alight upon, and carry out, a coordinated plan—through simulating the process of agreeing on a plan based on the participants' common ground. The variety of philosophical analyses typically have a complementary goal: to analyze what it *means* to have a shared plan, rather than resolving the problem of how such plans are selected. Indeed, the problem is often sidestepped by assuming that a common plan is articulated linguistically, e.g., "let's move the table" or "we're going to the store." Indeed, Tuomela (2005) explains his account metaphorically by imagining a publicly visible plan (e.g., of roles and responsibilities in cleaning a park), to which participants can equally publicly "sign up."

It is clear, though, that much coordinated action occurs with little or no explicit verbal interaction. Even where verbal agreements are formulated, these are typically incomplete. Thus, when two people walk to the coffee shop together, there is no explicit agreement to go at the same pace, and not to hop, walk backwards, stop unexpectedly, make a long phone call, and so on. But these "ground-rules" and many more are nonetheless implicitly agreed, as is evident by the fact that to violate one of them would require an explanation or apology (e.g., Garfinkel, 1967).[21]

Pluralistic ignorance. Building on our previous example (in which an agent believes the others are anarchists, where none in fact are), interesting implications emerge from the fact that the virtual bargain depends only on common ground for the phenomenon of pluralistic ignorance (e.g., Prentice & Miller, 1993). Thus, everybody at a party may believe that it is in common ground that students at a particular university like to binge drink, and disapprove of anyone who does not, even though each party may actually disapprove of binge drinking and privately envy those who avoid it. This has the perverse consequence that each individual may subscribe to the presumed result of the virtual bargain and affirm that "we plan to drink heavily tonight" even

[21] Virtual bargaining may help reconcile a dispute in the philosophy of shared intentions. According to Bratman (e.g., 1992, 1993), a joint commitment or shared intention requires each member of a group to be personally committing to carrying out their part in a joint plan. According to Gilbert (1987, 2006), it requires instead the group to be committed "as a body," which cannot be reduced to commitments of individual members. The plot of a G. K. Chesterton (1908) novel illustrates how these views appear to diverge. A group of seven putative anarchists plot to overthrow the state, not realizing that none are genuine anarchists (indeed, six are undercover detectives). Gilbert's account allows that the group intended, as a body, to overthrow the state. Bratman's account denies this because its individual members do not have this intention.

The virtual bargaining account helps capture both perspectives. Given their common ground (crucially not including each individual's private knowledge of being a detective), each putative anarchist believes that the virtual bargain is to individually and collectively overthrow the state and behaves in accordance with the bargain. Moreover, as Gilbert notes, each member of the group will anticipate, and probably receive, sanction from the others if they fail to "play their part." Thus, it is the existence of this virtual bargain that makes the group committed to the plot "as a body." But each individual has private knowledge (that they are detectives) which at some point will cause them to betray the plot. Thus, in line with Bratman's account, the plot is in a sense entirely illusory, because no individual will ultimately carry it through. Thus, Gilbert's and Bratman's perspectives appear compatible rather than contradictory.

though each person would rather not—indeed, it might be the case that were the students to have a discussion, they might discover that they would all prefer to play table tennis or watch a movie. In these circumstances, virtual bargaining would have failed correctly to simulate the outcome of real bargaining—because people are unaware of each other's true preferences. Yet there is also a second possibility: that virtual bargaining might *correctly* predict that actual discussion would favor heavy drinking over a movie, even though this might violate each person's individual preferences, because contributions to the discussion would be shaped to conform with the expected consensus. That is, in some circumstances, real bargaining would lead to the perverse outcome, because people would be unwilling to reveal their private preferences through pressure to conform; and if people correctly mentally simulate this through virtual bargaining, virtual bargaining will have the same perverse result. Thus, virtual bargaining can become a self-enforcing mechanism in contexts of pluralistic ignorance: simulating the presumed agreement suppresses 'honest' expression of preferences, and (i) makes that agreement likely were overt discussion to occur; (ii) makes overt discussion less likely, as the result is presumed to be already known. Indeed, we suggest that virtual bargaining may be crucial to the stability of a variety of group behaviors that are out of line with individual preferences. Few of us enjoy being on the losing side of a debate; so, mentally simulating the outcome of such a debate may cause us to stifle our dissent, so that the debate never takes place at all.

The nature and origin of common ground. As we have noted, virtual bargaining and we-reasoning more broadly between two or more people can only proceed on the basis of their common ground. This raises the questions of what common ground can be assumed, how information enters common ground, and how people are able to establish what is in common ground when uncertainty prevails.

The now-familiar problem of infinite regress rules out the possibility of inferring common ground from private knowledge. Thus, to adapt an example from Clark (1996), if *A* privately knows that *B* is from New Zealand, then *A* will likely assume that *B* knows the name of the current Prime Minister of New Zealand (among many other things). But this conclusion is not in common ground. Indeed, *B* has no way of knowing that *A* knows it, because *A*'s premises are private. But if *B* simply tells *A* that she is from New Zealand, then this information is put in common ground, and the conclusions may be in common ground also.

The requirement that conclusions in common ground can only be based on premises in common ground has surprisingly strong implications. A notable feature of everyday world knowledge is that it is a densely interconnected and interdependent network (e.g., Fodor, 1983; Quine & Ullian, 1978), such that typical inferences (e.g., from being a New Zealander, to almost certainly knowing the name of the current Prime Minister of New Zealand) depends on background knowledge about what New Zealand is and what a Prime Minister is, which themselves depend on further knowledge about geography, politics, and much more. Indeed, each piece of knowledge depends on further background knowledge, and so on, without end.[22]

Thus, to avoid infinite regress, entire *systems* of background beliefs must therefore be treated as common ground by all parties. So, rather than attempting to build up common ground through individual knowledge, it seems unavoidable that general background knowledge, at least, is taken for granted as common ground. From this point of view, common ground may, perhaps counterintuitively, be more psychologically basic than individual knowledge. That is, it

[22] This observation has been called the "fractal" character of everyday reasoning (Oaksford & Chater, 1998), as each part of the justification for, or explanation of, a belief is as complex as the belief itself. This line of thinking has varied intellectual roots: to coherentist, rather than foundationalist, theories in epistemology (Olsson, 2021); to the origin of the frame problem in artificial intelligence (Pylyshyn, 1987); and to the defeasible nature of human reasoning (Oaksford & Chater, 2007).

may require active cognitive processing for a person to recognize that information that they individually know is not in common ground, and to take account of this fact. Such processing is often effortful and not fully carried out. Indeed, this provides a new perspective on the so-called "curse of knowledge" (Camerer, Loewenstein, & Weber, 1989) in which people consistently assume that others know what they themselves know. The curse of knowledge can lead to poor outcomes in economic bargaining (Camerer et al., 1989); to difficulties in reasoning about false beliefs for both children and adults (Birch & Bloom, 2007); and to a tendency for the sender of a message to underestimate how difficult the receiver may find it to decode (Newton, 1990). Specifically, we suggest that the curse of knowledge should be viewed as following from the unavoidable default assumption that one's knowledge is in common ground.

The ability to communicate and interact successfully with others will depend, of course, on how far these assumptions of common ground are correct. If people's assumptions about common ground are largely aligned in a specific context, then they will interpret each other's behavior in the same way, and social interaction is likely to proceed smoothly. The importance of shared common ground in coordinating social behavior may help explain the strong positive affect associated with perceiving that we have a shared perspective with others, and the strong negative affect when we do not. Crucially, as Shteynberg et al. (2020) point out, this positive affect is generated not merely by the shared common ground, but by shared attentional focus being directed on the existence and relevance of this common ground. Similarly, the subjective experience of I-sharing—the momentary subjective experience of sharing a sense of self with another person—may reflect awareness of common ground assumptions being strengthened (e.g., if we face a common challenge or trauma, mimic each other's behavior, sing or dance in synchrony, and so on, Pinel, Long, Landau, Alexander, & Pyszczynski, 2006). The subjective

perception of such cognitive alignment with others is one way to understand the important concept of "shared reality" (Echterhoff, Higgins, & Levine, 2009; Higgins, 2019; Rossignac-Milon, Bolger, Zee, Boothby, & Higgins, 2021).

Perhaps even more basically, the drive to establish common ground—and hence facilitate future virtual bargaining that can coordinate social behavior—may provide a motivation for the enormous amount of time people spend in apparent "idle" conversation. Thus, although some anthropologists have assumed gossip to be an analog of grooming in non-human apes (e.g., Dunbar, 1998), it may be more appropriate to see conversation as establishing and solidifying the common ground for coordinating future behavior. This point is applicable particularly to what Clark (1996) calls *personal* common ground, which refers not to general background knowledge shared by the community at large, but to information shared by specific pairs or small groups of individuals. Building common ground through conversation and shared experiences seems to be a crucial part of building relationships, whether in friendships, romantic relationships, business partnerships, sports teams, or groups of workers.

Affiliation and groups. If social interactions are coordinated by tacit agreements, then the degree to which we will wish to enter such interactions will depend on our expectations about how easily tacit agreements can be formed, and how far they will be followed rather than flouted. The ability to form such agreements, as we have seen, will depend crucially on common ground—so that interactions will, other things being equal, proceed more smoothly for people with similar backgrounds, training, experiences, and so on. Thus, in a telecoms company, managers with a background in sales may find it easier to form tacit agreements with each other than with the engineers, and vice versa. This may lead to increased levels of liking, friendship, and further interactions within rather than between groups, potentially reinforcing the differences

in common ground between sales people and the engineers. On the other hand, lack of diversity may reduce the range of knowledge and experience available and hence reduce the quality of the resulting decisions.

A further interesting topic for future research is whether such effects are amplified by attribution errors arising from the curse of knowledge. If people in different groups tend to overestimate their common ground, then failed interactions are more likely to be attributed to "bad faith" on the part of the members of the other group. For example, if an engineer gives a technical explanation that is incomprehensible to the manager, this might arise because the engineer wrongly assumes this relevant technical knowledge as common ground. But it might be interpreted as a deliberate attempt to show off, obfuscate, or otherwise be unhelpful. This may be one of many forces maintaining in-groups and driving them away from out-groups, and more broadly causing people to affiliate with people they perceive to be like themselves. It might also suggest that contact between members of different groups may reduce such effects, especially if such contact leads to successful social coordination and collaboration (Paolini, Harwood, Hewstone, & Neumann, 2018). Indeed, recent experimental work suggests that even shared attention (e.g., to a brief video arguing for or against evolution by natural selection) may selectively increase liking for another person (Haj-Mohamadi, Fles, & Shteynberg, 2018). Perhaps tellingly, this effect only occurs if the shared information is congruent with a person's beliefs (i.e., whether the person believes in natural selection or creationism). Where the new information is not congruent with their prior assumptions, the person is blocked from encoding it as common ground between them and their partner, because they do not believe it themselves.

Relation to social norms. As we have seen, flexibility is a hallmark of human social interaction. From this standpoint, it seems misleading to think of the social life as guided

predominantly by "norms," viewed as fixed rules of behavior like the rules of chess.[23] Very subtle changes in the environment, or how a person acts or reacts, can completely change a social situation, depending on the underlying implicit agreements in play. Suppose two people approach a supermarket check-out at around the same time. If one person slows and waves to the other to go first, this may be interpreted as politely and helpfully breaking the deadlock. But suppose that a faster moving young person dashes slightly ahead of a slower moving older person. Now, if the older person slows and waves the other through, this may signal exasperation, or even hostility. The act of signaling draws attention to the deadlock, and that the younger person has "resolved" it, without agreement, to their own advantage. Slight exaggerations of the stopping and waving may intensify the communication of frustration. The younger person then faces the difficult question of whether to acknowledge the gesture, as if it were meant sincerely; to ignore it; or perhaps to apologize and let the older person go first. Such interactions are better viewed as complex, ad hoc, improvisations, rather than the playing out of a game according to fixed rules (Bertinetto & Bertram, 2020).

Still, though, many interactions are repeated, so that particular patterns of behavior will become increasingly conventionalized (although any conventions will always be open to challenge, irony, exaggeration, subversion, and so on). Indeed, it is interesting to speculate that social norms may best be viewed as emergent patterns arising over time from specific improvised interactions, by analogy with the way in which many linguists and psychologists in the "usage-based" tradition see grammatical patterns as arising through gradual entrenchment of, and generalization from, specific uses (Hopper & Traugott, 1993; Tomasello, 2003). This viewpoint has the implication that virtual bargaining may be essential to the development of

[23] Although this may sometimes be a useful theoretical approximation (Hawkins, Goodman, & Goldstone, 2019).

human communicative and social norms, and thus that species without the ability to engage in virtual bargaining would be unable to develop a complex culture.

The importance of scripts, roles, and responsibilities. Many social situations (greetings, being seated or served by a waiter, going through a religious ritual, receiving a certificate at a school assembly) are governed by something close to standardized "scripts" (Schank & Abelson, 1977), in which different participants have distinct roles and responsibilities. For example, a waiter is expected to ask a customer for their order, not the reverse; the waiter is responsible for passing the order to the chef, and so on. Where scripts are in common ground, they can, of course, substantially simplify the problem of social coordination—but a coordinated social interaction requires continual negotiation (and, often, repair) for the script to be "performed" successfully. The complexity of such coordination requires that most such negotiation cannot be mediated by language, but must be predominantly "virtual;" and the need for different parties to arrive at the same solution, out of multiple possibilities, raises, as ever, the paradox of social interaction: that each party is attempting to predict, and coordinate with, the actions of the other. So, for example, both the waiter and customer need to agree which table the waiter is attending to (the waiter may pass near a table, for example, merely to allow other customers to pass). They need to agree, too, when it is appropriate to request, or to place, and order (e.g., not if the waiter approaches the table with a dustpan and brush, after glass has been broken). Indeed, the virtual bargaining viewpoint meshes with a dramaturgical perspective on social behavior, as the joint creation of a shared "performance" (Goffman, 1959).

Deciding who "we" are. We-reasoning, including virtual bargaining, begins from presumed common ground about who "we" are—that is, who the parties are in the social interaction. Simple physical presence is not, of course, sufficient. Thus, if one diner suggests

going to a movie to another, a waiter who is clearing the plates, or a diner at the next table, would not be expected to be included in the invitation. Thus, an important part of any social interaction is ensuring that there is common ground among participants that they *are* participants.[24]

Misconstruing who is, or is not, engaged in a particular social interaction can itself be a source of social confusion and embarrassment—as when we wave back at someone who was signaling to someone else; or when taking a seat that is actually being offered to someone else. An open theoretical question concerns how one's legitimate participation in a social interaction is itself commonly agreed. Assuming that this is itself achieved via virtual bargaining seems to raise the specter of circularity, because such bargaining itself presupposes an agreement about who is involved in the agreement. We suspect the circularity may be more apparent than real: that the question of who is *in* the virtual bargain is really part of the bargain itself (just as the terms of a business contract include the identities of the participating firms and individuals).

Nonetheless, the issues here are subtle. For example, consider informal and apparently tacit conventions, such as those described by Sugden (1989): in a Yorkshire fishing village, anyone could mark a pile of driftwood as theirs (for later collection) by placing two stones on it. A villager could not, presumably, violate the rule and take driftwood for themselves merely by declaring that they were not part of the "bargain." The normative force of the rule is that people in the relevant social group cannot simply declare themselves outside the bargain and hence not bound by it (even if they say that they do not object to others taking their driftwood). Relatedly, in some contexts, the question of who is party to a social interaction, and hence drawn into a virtual bargain, may be contentious: a fundraiser with a tin may wish to engage a hapless

[24] It need not be in common ground to non-participants that they are not included—indeed, non-participants to social interactions need not even be aware that the interaction is occurring.

shopper, creating mutual expectation of a donation to a good cause; the shopper may avoid eye contact and pretend not to notice the fundraiser's existence, to avoid "implicitly agreeing" to be part of any such bargain.

The question of who "we" are is likely to be closely linked to questions of social identity and group formation. A common identification that we are members of the same fishing community, regiment, religious group, political party, or nation may establish common ground that, in some ways at least, we are part of a common social unit, governed by an agreed social contract (e.g., to follow group norms, help each other, and so on). Thus, from this viewpoint, social identity (Tajfel, 1974) as being part of a group depends not just on self-identification as a group member, but may depend on group membership being in common ground within the group ("we all know who we are"). This common ground is required for group members to coordinate effectively through virtual bargaining. Indeed, work in the minimal group paradigm has long shown that merely establishing in people's common ground that they have been divided into two arbitrary groups (e.g., randomly given red versus blue t-shirts) leads to greater cooperation within rather than between groups (e.g., Brewer, 1979). This is often interpreted as indicating greater liking for, and hence pro-social behavior towards, in-group than out-group members. According to the present perspective, this same manipulation might also be expected to differentiate performance even in so-called Schelling games (Schelling, 1960), where people are each rewarded if they independently make the same choice as each other from a number of options (e.g., colors, dates, numbers, locations). As we have noted, in coordination problems of this kind, enhanced pro-social motivation does not improve performance. What is crucial is the ability to imagine what "we" would agree were we able to discuss what would be most natural (e.g., "choose the most common color," "choose January 1st," etc.). By priming or suppressing

the "we" perspective through having people interact who are either in the same, or a different, arbitrary group, we would expect such inferences, and hence resulting coordination, to be enhanced or impeded. As far as we know, this prediction remains to be tested experimentally.

A new perspective on reciprocation. Many mutually beneficial social and economic interactions unfold over time, in which *A* helps *B* on some occasions, and *B* helps *A* on others. Indeed, some patterns of mutual help and the common expectation of such help is often a foundation for mutually beneficial relationships, rather than momentary transactions, from friendships, to life-partnerships, to business partnerships, and alliances of all kinds. Patterns of reciprocation are puzzling for traditional reason-based models of behavior. To see why, consider the well-known "Centipede" game.[25] The game has a fixed maximum number of alternating turns (e.g., 10) between two players. On each turn, the player whose turn it is to move can either (a) give up $1 of her own money, so that the other player receives $3; or (b) stop the game. Over time, both players benefit from reciprocal interaction: each gives away a number of $1 sums, and receives a roughly equal number of $3 sums. But if the game has a commonly known endpoint (e.g., each player can play just 10 trials), then a standard rational account is that no reciprocation is possible. Both players anticipate that the player with the final "turn" (say, *A*) will not send back a dollar (having nothing more to gain, because the other cannot reciprocate). They can then infer that *B* will not send back the dollar on the penultimate turn (knowing that *A* will not reciprocate on the final turn). Following this logic to the start of the game through "backward induction," leads to the conclusion that reciprocation can never start. Or, to consider a related example, consider a finitely repeated Prisoner's Dilemma (PD) (Embrey, Frechette, & Yuksel, 2018), in which each player can cooperate or defect on each turn (a simple version of Prisoner's

[25] The Centipede game was originated in Rosenthal (1981). The version we describe here is Rosenthal's original version, but many variations have since been developed and explored experimentally.

Dilemma is that, on each turn, both players simultaneously choose whether to give up $1 of their own money, so that the other player receives $3). On a single round of PD, of course, it is in each player's interest to withhold their dollar (whatever the other does, this will leave them one dollar better off). But if PD is repeated, say, twenty times, then there is a substantial opportunity for mutual benefit by reciprocally cooperating. As in the Centipede game, rational agents seem unable to obtain such benefits, because of the logic of "backward induction:" both know that defection will occur on the last trial; so it is rational to defect on the second to last trial, and the third to last, all the way to the beginning of the game. So, the apparently substantial opportunity for reciprocation (I'll help you, if you help me) appears inaccessible. In practice, people typically do engage in high levels of cooperation in both the Centipede and finitely repeated PD games, though sometimes ceasing cooperation near the very end of the game (Embrey et al., 2018; García-Pola, Iriberri, & Kovárík, 2018).[26]

The virtual bargaining viewpoint outlined here suggests that reciprocation need not be treated as a distinct phenomenon (e.g., Fehr & Gächter, 2000b), but rather viewed as a special case of reaching a virtual agreement. Specifically, virtual bargaining can tell the players that, were they able to communicate, they would agree to be cooperative—i.e., to keep paying across sums of $1 until the end of the game. While this hypothetical agreement is in common ground, each player may reasonably wonder: "what happens if I stick to the agreement, but the other "exploits" my doing so to their own best advantage?" In the Centipede and finitely repeated PD

[26] The picture is different where the PD is repeated indefinitely; then, the so-called "folk theorem" of game theory applies, which states that (typically) a wide variety of outcomes is rationally justifiable (e.g., Gintis, 2000)—but the players then face the coordination problem of simultaneously alighting on the same one, which has been the primary topic here. Thus, the need for virtual bargaining, or some similar mechanism, recurs in a rational account of infinitely repeated PDs. With one-shot PDs, virtual bargaining does not necessarily generate a cooperative outcome. By contrast, with the finitely repeated PD, a virtual bargain to cooperate throughout is credible, because even if one agent follows the agreement and the other best-responds, this implies that defection will only occur on the last turn, so that both agents can accrue significant benefits for prior cooperation.

games, the best way to exploit someone who dutifully follows the virtual agreement is to continually cooperate until the very last turn. By exploiting the dutiful player only at the last moment, the selfish player gains all the accumulated benefits of mutual cooperation, and thus maximizes their own payoff. Crucially, however, the payoff to the dutiful player is still high, because they, too, gain the accumulated benefits of mutual cooperation, except when exploited on the very last turn. So, the players can anticipate a good outcome for themselves from the agreement to cooperate to the end of the game, even where they suspect that the other might secretly intend to act selfishly. Each player can reason, "Of course, I can trust myself to stick to a virtual agreement to cooperate throughout the game; and even if the other player selfishly exploits me, my outcome is still good. So, this is a good agreement for me, whether the other person sticks to the agreement or behaves selfishly." And this reasoning is itself in common ground, so it is commonly known to both parties that the agreement is mutually advantageous and resilient against exploitation (for a formal analysis, see Melkonyan et al., 2021).

This perspective is very different from many "algorithmic" perspectives on reciprocation, according to which people (or artificial agents) follow rules, such as tit-for-tat, which are assumed to have "evolved" either through cultural or biological evolution (e.g., Axelrod & Hamilton, 1981 and the large subsequent literature). By contrast, the present viewpoint sees reciprocation as a paradigm example of a "social contract" which will be in common ground to both parties, with common expectations that the contract will be fulfilled (and the expectation of complaint to, or even punishment of, the other if it is not).

To see the difference, consider a (real) case in which *A* tidies *B*'s garden each week; and each week *B* gives *A* a bottle of wine. It turns out that *A* does not drink wine, and hence the "reciprocation" is entirely void, with the bottles simply accumulating undrunk. *A* nonetheless

feels obligated to continue tending *B*'s garden, rather than violating the unspoken agreement (although clearly an agreement has arisen due to a misunderstanding). Suppose *A* were to discover that *B* does not drink wine (and *B* were to know that *A* had discovered this). Then, a conventional rational analysis would lead to the conclusion that *A* could abruptly stop giving wine, and expect *B* to tend the garden as before (reasoning that *B* must enjoy the gardening for its own sake, as reciprocation is clearly not *B*'s motive). Yet this would almost certainly be viewed as extremely rude, and *B* might very well cease gardening at once. Instead, *A* would need to find some other (one hopes more successful) way of reciprocating—thus forming a new tacit "contract" with *B* (perhaps *B* could take produce or flowers from the garden). Or suppose *A* goes on a summer holiday for two weeks, and hence misses a "payment" of wine to *B*. In such novel circumstances, the two parties may not have sufficient basis for coordinating on the same virtual bargain, and hence disagreements may arise. Quite likely *B* will garden anyway, and anticipate two bottles of wine on *A*'s return. But *A* might provide just one bottle, assuming that a single bottle is a token of appreciation, rather than an implicit "payment" per week for *A*'s gardening. There would be at least the possibility that *B* might feel affronted if *A* did not provide some extra token of appreciation on her return, even if the token is entirely unwanted. Indeed, irritation and even anger are likely to be common responses when virtual bargaining fails.

These patterns of thought and behavior are intuitively natural, and make sense if both parties think they have entered a tacit agreement. But they do not follow from any algorithmic account of reciprocation, for example, following a rule such as tit-for-tat.[27] Furthermore, this

[27] In evolutionary game theory (Axelrod & Hamilton, 1981), algorithmic strategies such as tit-for-tat are primitive behavioral patterns which are selected for over repeated interactions between agents. Thus, the algorithms that predominate are those which have met with greatest success in the past. By contrast, in the virtual bargaining approach, strategies (which may include tit-for-tat) arise through a process of forward-looking choice, in view of their likely success in the current interaction.

viewpoint helps explain why people may *reject* help from another, when this is apparently against their interests—because accepting help may, in some contexts, signal tacit acceptance of a mutual bond of reciprocal support to which one party may be reluctant to sign up (as when one accepts an offer of help from the mafia, Gambetta, 1996).

Links with other types of contract-based theorizing in psychology. The present account of how "social contracts" are improvised in the moment has interesting parallels in other ideas in psychology. In the psychology of language, Brennan and Clark (1996) propose that "conceptual pacts" coordinate moment-by-moment communication (a conceptual pact might be: "let's call this odd-shaped tangram pattern 'the rocket'" or "in this conversation, 'Ali' refers to my friend rather than the legendary boxer"). More broadly, social scientists frequently talk of people negotiating meanings, identities, and relationships (e.g., Thomas, Sargent, & Hardy, 2011)—but of course much of this negotiation is not explicit, and may perhaps usefully be considered as stemming from virtual bargaining.

A second connection focuses on a putative psychological contract between employees and companies (Argyris, 1960; Rousseau, 1995; Rousseau & Shperling, 2003). Again, virtual bargaining provides a potential for understanding how such psychological contracts can become established without, or with a minimum of, explicit communication. We believe that it is crucial to understand such contracts as jointly recognized and understood. Thus, for example, if a worker is not treated as they expect by their employer (e.g., their hours and pay are reduced without prior notice), their outrage is, we suspect, not merely that the employer has deviated from the employee's understanding of their implicit agreement. It is rather that the employer has knowingly and willfully violated an implicit agreement to which they had both "signed up." The employee thus feels a sense not merely of having been maltreated, but betrayed. It may therefore

be useful to recast the important literature on psychological contracts from the viewpoint that psychological contracts are joint contracts (or at least are perceived as joint contracts), rather than being mentally represented by a single individual.

A third related theoretical idea is what Rousseau (1989) called "implied" contracts. These are implicit rules by which society at large guides and governs relationships (e.g., between employee and employer), whether or not these are mentally represented by either party. What is the status of such implied contracts? We suggest that they may be analogous to the rules governing language, which are not explicitly formulated by any individual speaker but are distributed through a linguistic community. The parallel with language also suggests that implied contracts, and implicit social norms more generally, may become established over time through the gradual accumulation of virtual bargains each of which deals with specific circumstances. Each new virtual bargain will build on those that have gone before, via processes of entrenchment and generalization, potentially producing an increasingly conventionalized, rule-like system.

Relation to moral reasoning. These considerations suggest, more broadly, that social interactions are governed by tacit agreements with a "moral" dimension. That is, a commitment to the "rightness" of the tacit agreement, having an obligation to uphold it, and accepting the appropriateness of chastising or punishing those who violate it without good cause, is *part* of committing to the agreement itself. Indeed, when people violate tacit agreements, even in low-stakes social interactions concerned with seating preferences, borrowing pens, or handing each other books, the reaction of others can range from mild irritation to outrage. More broadly, a model of social interaction founded on mutual tacit agreements, rather than second-guessing the thoughts and actions of another (who is second-guessing one's own), explains the inherently

normative nature of social behavior, and that social norms are typically viewed as common ground between participants.

From a traditional individualistic perspective, it is natural to think of people as focused primarily on the consequences of their own and other people's actions. From this point of view, we might expect people to praise and reward the actions of others that have positive consequences for them, and to punish the actions of those that have bad consequences for them (which might be the foundation of reciprocal cooperation, Fehr & Gächter, 2000a). But the examples above indicate that people are also greatly concerned with people following tacit (and indeed explicit) agreements. From this point of view, the human ability to engage in coordinated social behavior may be continuous with our sense of the "appropriate" way to behave (March & Olsen, 2008), a viewpoint that is consistent with developmental evidence (e.g., Tomasello, 2020). Indeed, if rules of local social interaction are continuous with wider moral thinking, then virtual bargaining may provide a starting point for understanding moral thinking more broadly, and especially those that naturally flow from contractarian approaches to ethics (e.g., Gauthier, 1986; Scanlon, 1998).

Virtual bargaining and verbal communication. We have argued that virtual bargaining provides a mechanism through which people can coordinate their behavior without communication. But even where verbal communication is possible, the paradox of social interaction, and the need to resolve it by virtual bargaining or a similar mechanism, still arises. Communicative signals—from facial expressions, to gestures, to complex symbolic language—are invariably highly ambiguous (Christiansen & Chater, 2022), and require complex pragmatic processes to determine a specific meaning (Clark, 1996; Grice, 1975; Levinson, 2000; Sperber & Wilson, 1986). Communication will only succeed, of course, where the speaker and hearer alight

on the *same* meaning. Thus, the pragmatic inference involved in interpreting a communicative signal is a paradigm example of a coordination problem. Suppose the speaker points vaguely and asks, "Can you pass me that one?" The speaker has to work out (among other things) what the hearer thinks "that one" is; and the hearer has to work out what the speaker thinks "that one" is. This problem of mutual prediction cannot be resolved by any amount of recursive mentalizing (Clark, 1996).

Although further communication may help resolve uncertainty (Healey, Mills, Eshghi, & Howes, 2018), any such communication will itself be ambiguous and need further interpretation. Virtual bargaining provides a possible route for breaking out of this regress: in the light of their common ground, both parties ask, "Given our objectives, what would be the most efficient way to map possible signals to meanings?" If they can jointly infer a signal-meaning mapping in the present communicative context, given their common ground, then this mapping can be applied by both parties to interpret whatever signal happens to be sent. For example, if the two people are doing some DIY together and have a pile of tools on the floor between them, "that one" must be a screwdriver, given their common ground that their current task is fixing a screw to the wall. If the speaker wanted the hammer right next to the screwdriver, they would need to be more specific ("that hammer"). The shorter, simpler phrase is reserved for the most salient interpretation, given their common ground, which follows standard efficiency considerations (e.g., Gibson et al., 2019). This approach is in the spirit of Clark's (1996) theory of communication, and has been recently developed using the virtual bargaining framework and related work (Bundy, Philalithis, & Li, 2021; Chater & Misyak, 2021; Stacy et al., 2021).

How much virtual bargaining is carried out 'in the moment'? An important direction for future work is to clarify whether and when virtual bargaining occurs in the moment, or whether it

is better considered as a "rational analysis" (e.g., Anderson, 1990; Chater & Oaksford, 1999) for repetitive behavior that is long-established through individual learning or cultural evolution. This issue arises, of course, throughout psychology: for example, in considering instance-based versus algorithmic theories of skill learning (e.g., Logan, 1988), construction vs rule-based views of syntax (e.g., Ziegler, Bencini, Goldberg, & Snedeker, 2019), Gricean-style natural language pragmatics (Goodman & Frank, 2016), or the interpretation of metaphor (Glucksberg, 2003). Such questions are, of course, typically difficult to answer in any domain. We suggest, though, that the flexibility of social interactions, and their responsiveness to very subtle communicative cues or shifts in background knowledge (see, e.g., the astonishing flexibility of the interpretation of newly encountered communicative signals, e.g., Clark, 1996; Galantucci, 2005; Misyak, Noguchi, & Chater, 2016), suggests that a good deal of processing must be improvised in each fresh communicative interaction. Nonetheless, as with skill learning more broadly, it is likely that each new social interaction will be shaped by prior experience of similar interactions, gradually leading to the "entrenchment" of certain patterns of behavior. Thus, as noted above, systems of increasingly rule-like patterns in behavior would be expected to arise both during learning and development within the individual, and within newly formed social groups. At each point, though, social behavior seems highly malleable and adaptable to the demands of the moment, which seems to defy codification and require in-the-moment adjustment by social actors.

Conclusions

The shared intentionality of social interaction, and the mutual dependence of the minds, is ubiquitous and relatively uncontroversial. Yet it appears to have paradoxical consequences for

many theories of the psychological foundations of social behavior. In particular, a theoretical circularity seems to arise where *A*'s thoughts and actions depend on *A*'s interpretation of *B*'s thoughts and actions; yet *B*'s thoughts and actions depend on *B*'s interpretation of *A*'s thoughts and actions. Even formulating this apparent problem of circularity leaves one feeling rather befuddled. It seems initially possible that this might be no more than a pseudo-problem, somehow generated by loose argument or use of language; or perhaps a genuine conceptual puzzle, but one with few implications for the psychologist. We suggest the opposite: that the mutual prediction that is essential to social interactions generates a paradox that psychological theories ignore at their peril. Thus, theoretical accounts that see social interaction as involving each person predicting the other's behavior (Tamir & Thornton, 2018), reading their mental states (Gopnik & Wellman, 1992), or forming first-, second-, or higher-order beliefs about others (e.g., Liddle & Nettle, 2006) will inevitably be incomplete. More generally, theories that aim to explain social behavior in terms of reasons are coherent only to the extent that the reasons that they provide are rationally coherent—and hence free of paradoxes. We have seen that in the context of reciprocal social interactions avoiding paradox is surprisingly difficult.

In this paper, we have considered simple social interactions in which the paradox arises very starkly, even though the "natural" solution is highly intuitive. Thus, social interaction seems profoundly different from either social perception (interpreting the behavior of others through observation) or social action (one-way influence on the behavior of others). This observation has, we believe, an important implication for the empirical study of social behavior—that by abstracting away from *interaction*, as has been common in social psychology for many decades, psychologist are missing a crucial aspect of social cognition and behavior.

The solution to the paradox of social interaction requires, we have argued, thinking about what "we" should do—a shift from I-thinking to we-thinking. And the virtual bargaining account aims to explain we-thinking without invoking the shadowy notion of a group mind. Instead, we propose that individuals can we-reason successfully by asking: what would we agree, if we can discussion and bargain? Where the result of such simulated virtual bargaining is self-evident to all parties, then results of such bargaining can be followed directly. Understanding the nature and limits of virtual bargaining, how people decide who is part of the bargain, how the common ground underpinning virtual bargaining is established, and how the potential for successful virtual bargaining affects whether people wish to join or maintain relations, or group membership, remains interesting questions for future research. As we have seen, virtual bargaining provides a possible mechanism for the cumulative creation and entrenchment of linguistic conventions and social norms. Moreover, the normative force of virtual bargaining—that participants in the bargain feel that they and others ought to follow it—raises interesting connections with moral psychology.

The argument of this paper has implications for how computational models of non-social behavior in the cognitive and brain sciences may extend to social interaction. Computational theories of human cognition, rooted in cognitive science, machine learning, artificial intelligence, and neuroscience, have become increasingly sophisticated and successful. Yet, we argue, these models are, in a specific and we suggest important sense, asocial. Specifically, the mutual prediction in social interaction—the fact that each participant is attempting to predict, respond to, and accommodate the thoughts and actions of the other—lies outside the scope of such models. Indeed, the attempt to apply standard models of human cognition to mutual social interaction leads, as we have seen, to paradoxical results.

According to the argument of this paper, the process of virtual bargaining is fundamental to the human ability to engage in complex social interactions. Yet we are often oblivious to this process of bargaining in part, we suggest, because the highly sophisticated social reasoning underpinning our interactions is so familiar and natural to us. Just as we have the illusion of "direct" contact with the visual environment, unaware of the spectacularly complex calculations carried out by our visual brains (e.g., Ullman, 1980), so the world of social interaction seems "transparent" to us, even though rich and subtle background calculations are in play.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Akerlof, R. (2016). "We thinking" and its consequences. *American Economic Review, 106*(5), 415-419. doi:10.1257/aer.p20161040

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Argyris, C. (1960). *Understanding Organizational Behavior*. Homewood, IL: The Dorsey Press.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs, 70*, 1-70.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics, 4*(6), 1236-1239. doi:10.1214/aos/1176343654

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396. doi:10.1126/science.7466396

Bacharach, M., Gold, N., & Sugden, R. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton, NJ: Princeton University Press.

Baker, C. L. (2012). *Bayesian Theory of Mind: Modeling Human Reasoning about Beliefs, Desires, Goals, and Social Relations*. PhD Thesis: Department of Brain and Cognitive Sciences, MIT.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4). doi:10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329-349. doi:10.1016/j.cognition.2009.07.005

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*(2), 122-147. doi:10.1037//0003-066x.37.2.122

Bandura, A., Ross, S. A., & Ross, D. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63*(3), 575-582. doi:10.1037/h0045925

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241-251. doi:10.1111/1469-7610.00715

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica, 52*(4), 1007-1028. doi:10.2307/1911196

Bertinetto, A., & Bertram, G. W. (2020). We make up the rules as we go along: Improvisation as an essential aspect of human practices? *Open Philosophy, 3*(1), 202-221. doi:10.1515/opphil-2020-0012

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics, 4*(2-3), 174-200.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*(5), 382-386. doi:10.1111/j.1467-9280.2007.01909.x

Boesch, C. (1994). Cooperative hunting in wild chimpanzees. *Animal Behaviour, 48*, 653-667.

Bohn, M., & Koymen, B. (2018). Common ground and development. *Child Development Perspectives, 12*(2), 104-108. doi:10.1111/cdep.12269

Boothby, E. J., Clark, M. S., & Bargh, J. A. (2014). Shared experiences are amplified. *Psychological Science, 25*(12), 2209-2216. doi:10.1177/0956797614551162

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences, 108*, 10918-10925. doi:10.1073/pnas.1100290108

Bratman, M. E. (1992). Shared cooperative activity. *Philosophical Review, 101*(2), 327-340. doi:10.2307/2185537

Bratman, M. E. (1993). Shared intention. *Ethics, 104*(1), 97-113. doi:10.1086/293577

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482-1493. doi:10.1037/0278-7393.22.6.1482

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*(2), 307-324. doi:10.1037/0033-2909.86.2.307

Bullinger, A. F., Wyman, E., Melis, A. P., & Tomasello, M. (2011). Coordination of chimpanzees (Pan troglodytes) in a stag hunt game. *International Journal of Primatology, 32*(6), 1296-1310. doi:10.1007/s10764-011-9546-3

Bundy, A., Philalithis, E., & Li, X. (2021). Modelling virtual bargaining using logical representation change. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence*. Oxford, UK: Oxford University Press.

Byrne, R., & Whiten, A. (1989). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford, UK: Oxford University Press.

Byrne, R. M. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Byrne, R. W., & Whiten, A. (1990). Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. *Behavior and Philosophy, 18*(1), 73–75.

Call, J. (2009). Contrasting the social cognition of humans and nonhuman apes: The shared intentionality hypothesis. *Topics in Cognitive Science, 1*(2), 368-379. doi:10.1111/j.1756-8765.2009.01025.x

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*(5), 1232-1254. doi:10.1086/261651

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics, 119*(3), 861-898. doi:10.1162/0033553041502225

Chater, N., & Misyak, J. (2021). Spontaneous Communicative Conventions through Virtual Bargaining. In S. Muggleton & N. Chater (Eds.), *Human-Like Machine Intelligence* (pp. 52-67). Oxford, UK: Oxford University Press.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*(2), 57-65. doi:10.1016/s1364-6613(98)01273-x

Chesterton, G. K. (1908). *The Man Who Was Thursday: A Nightmare*. London: J. W. Arrowsmith.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Boston, MA: Little, Brown Spark.

Christiansen, M. H., & Chater, N. (2022). *The Language Game*. New York, NY / London, UK: Basic Books / Bantam Books.

Cialdini, R. B. (2001). *Influence: Science and Practice (4th ed.)*. Boston: Allyn & Bacon.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591-621. doi:10.1146/annurev.psych.55.090902.142015

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. doi:10.1017/s0140525x12000477

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H. (2021). Anchoring utterances. *Topics in Cognitive Science, In press: https://onlinelibrary.wiley.com/doi/full/10.1111/tops.12496*. doi:10.1111/tops.12496

Clark, H. H., & Brennan, S. A. (1991). Grounding in Communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington: APA Books.

Collingwood, R. G. (1947). *The New Leviathan*. Oxford, UK: Oxford University Press.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences, 26*(2), 139-198. doi:10.1017/s0140525x03000050

Colman, A. M., & Gold, N. (2017). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review, available on: https://link.springer.com/article/10.3758%2Fs13423-017-1399-0*. doi:10.3758/s13423-017-1399-0

Colman, A. M., Pulford, B. D., & Rose, J. (2008). Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica, 128*(2), 387-397. doi:10.1016/j.actpsy.2007.08.003

Costa-Gomes, M. A., & Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review, 96*(5), 1737-1768. doi:10.1257/aer.96.5.1737

Couzin, I. D., Krause, J., Franks, N. R., & Levin, S. A. (2005). Effective leadership and decision-making in animal groups on the move. *Nature, 433*(7025), 513-516. doi:10.1038/nature03236

Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., . . . Brass, M. (2018). Automatic imitation: A meta-analysis. *Psychological Bulletin, 144*(5), 453-500. doi:10.1037/bul0000143

Cubitt, R. P., & Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of David Lewis' game theory. *Economics and Philosophy, 19*(2), 175-210. doi:10.1017/s0266267103001123

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy, 60*(23), 685-700. doi:10.2307/2023177

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (pp. 282-309). Oxford: Blackwell.

De Freitas, J., Thomas, K., DeScioli, P., & Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences of the United States of America, 116*(28), 13751-13758. doi:10.1073/pnas.1905518116

de Waal, F. (2007). *Chimpanzee Politics: Power and Sex Among Apes*. Baltimore: Johns Hopkins University Press.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Diamond, P. A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *Journal of Political Economy, 75*(5), 765-766. doi:10.1086/259353

Dunbar, R. I. M. (1998). *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard University Press.

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared reality experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science, 4*(5), 496-521. doi:10.1111/j.1745-6924.2009.01161.x

Elster, J., & Roemer, J. E. (Eds.). (1993). *Interpersonal Comparisons of Well-Being*. Cambridge, UK: Cambridge University Press.

Embrey, M., Frechette, G. R., & Yuksel, S. (2018). Cooperation in the finitely repeated Prisoner's Dilemma. *Quarterly Journal of Economics, 133*(1), 509-551. doi:10.1093/qje/qjx033

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature, 414*(6862), 443-446. doi:10.1038/35106560

Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980-994. doi:10.1257/aer.90.4.980

Fehr, E., & Gächter, S. (2000b). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives, 14*(3), 159-181. doi:10.1257/jep.14.3.159

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmuller, A., Frey, D., . . . Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517-537. doi:10.1037/a0023304

Fishbein, M. (1979). A theory of reasoned action: Some applications and implications. In H. E. Howe & M. M. Page (Eds.), *Nebraska Symposium on Motivation* (Vol. 27, pp. 65-116). Lincoln: University of Nebraska Press.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science, 336*(6084), 998-998. doi:10.1126/science.1218633

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science, 29*(5), 737-767. doi:10.1207/s15516709cog0000_34

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences, 17*(4), 160-165. doi:10.1016/j.tics.2013.02.002

Gambetta, D. (1996). *The Sicilian Mafia: The Business of Private Protection*. Cambridge, MA: Harvard University Press.

García-Pola, B., Iriberri, N., & Kovárík, J. (2018). Non-equilibrium play in Centipede games. *Centre for Economic Policy Research (CEPR), Discussion Paper, DP11477*.

Garfinkel, H. (1967). *Studies in ethnomethodology*. New Jersey: Prentice-Hall.

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.

Georganas, S., Healy, P. J., & Weber, R. A. (2015). On the persistence of strategic sophistication. *Journal of Economic Theory, 159*, 369-400. doi:10.1016/j.jet.2015.07.012

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415*(6873), 755. Retrieved from <Go to ISI>://WOS:000173833900035

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165-193. doi:10.1016/0010-0277(95)00661-h

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389-407. doi:10.1016/j.tics.2019.02.003

Gilbert, M. (1987). Modeling collective belief. *Synthese, 73*(1), 185-204. doi:10.1007/bf00485446

Gilbert, M. (1989). *On Social Facts*. New York: Routledge.

Gilbert, M. (2006). Rationality in collective action. *Philosophy of the Social Sciences, 36*(1), 3-17. doi:10.1177/0048393105284167

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies, 144*(1), 167-187. doi:10.1007/s11098-009-9372-z

Gintis, H. (2000). *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton, NJ: Princeton University Press.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences, 7*(2), 92-96. doi:10.1016/s1364-6613(02)00040-2

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818-829. doi:10.1016/j.tics.2016.08.005

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language, 7*(1-2), 145-171.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language, 1*(2), 158-171.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. In S. T. Fiske (Ed.), *Annual Review of Psychology* (Vol. 71, pp. 419-445).

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics: Speech Acts* (Vol. 3, pp. 41-58). New York: Academic Press.

Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the childs internalization of values: A reconceptualization of current points of view. *Developmental Psychology, 30*(1), 4-19. doi:10.1037//0012-1649.30.1.4

Grzyb, T., Dolinski, D., Trojanowski, J., & Bar-Tal, Y. (2018). Cognitive structuring and obedience toward authority. *Personality and Individual Differences, 133*, 115-120. doi:10.1016/j.paid.2017.08.032

Guinote, A., & Vescio, T. K. (2010). *The Social Psychology of Power*. New York, NY: Guilford Press.

Habermas, J. (1987). *The Theory of Communicative Action*. Boston: Beacon Press.

Haj-Mohamadi, P., Fles, E. H., & Shteynberg, G. (2018). When can shared attention increase affiliation? On the bonding effects of co-experienced belief affirmation. *Journal of Experimental Social Psychology, 75*, 103-106. doi:10.1016/j.jesp.2017.11.007

Hakli, R., Miller, K., & Tuomela, R. (2010). Two kinds of we-reasoning. *Economics and Philosophy, 26*(3), 291-320. doi:10.1017/s0266267110000386

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557-559. doi:10.1038/nature06288

Haney, C., Banks, C., & Zimbardo, P. (1973). Study of prisoners and guards in a simulated prison. *Naval Research Reviews, 26*(9), 1-17. Retrieved from <Go to ISI>://WOS:A1973R421500001

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*, 139-151. doi:10.1006/anbe.2000.1518

Harman, G. (1977). Review of "Linguistic Behavior" by Jonathan Bennett. *Language, 53*(2), 417-424.

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Hart, H. L. A. (1994). *The Concept of Law*. Oxford: Clarendon Press.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science, 2*(3), 96-99.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 512-521. doi:10.1016/s0022-5371(74)80003-4

Hawkins, R. X. D., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences, 23*(2), 158-169. doi:10.1016/j.tics.2018.11.003

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running repairs: Coordinating meaning in dialogue. *Topics in Cognitive Science, 10*(2), 367-388. doi:10.1111/tops.12336

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology, 81*(6), 1028-1041. doi:10.1037//0022-3514.81.6.1028

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Heyes, C. M., & Galef Jr, B. G. (Eds.). (1996). *Social Learning and Imitation: The Roots of Culture*. New York: Academic Press.

Higgins, E. T. (2019). *Shared Reality: What Makes Us Strong and Tears Us Apart*. New York, NY: Oxford University Press.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic books.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization*. Cambridge, UK: Cambridge University Press.

Hoppitt, W. J. E., Brown, G. R., Kendal, R., Rendell, L., Thornton, A., Webster, M. M., & Laland, K. N. (2008). Lessons from animal teaching. *Trends in Ecology and Evolution, 23*(9), 486-493. doi:10.1016/j.tree.2008.05.008

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, UK: Cambridge University Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136-153. doi:10.1037/0033-295x.93.2.136

Kalai, E., & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica, 43*(3), 513-518.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192-238). Lincoln: University of Nebraska Press.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). *Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction*. Proceedings of the 38th Annual Conference of the Cognitive Science Society.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT press.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*(731). doi:10.3389/fpsyg.2015.00731

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology, 4*(3-4), 231-244.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492-527. doi:10.1037/0033-295x.95.4.492

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*, 93-120. doi:10.1146/annurev.psych.51.1.93

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature Communications, 9*(1), 1-9. doi:10.1038/s41467-018-04925-y

March, J. G., & Olsen, J. P. (2008). The logic of appropriateness. In R. E. Goodin, M. Moran, & M. Rein (Eds.), *The Oxford Handbook of Public Policy*. Oxford: Oxford University Press.

Melkonyan, T., Zeitoun, H., & Chater, N. (2018). Collusion in Bertrand versus Cournot competition: A virtual bargaining approach. *Management Science, 64*(12), 5599-5609.

Melkonyan, T., Zeitoun, H., & Chater, N. (2021). The cognitive foundations of tacit commitments. *Working Paper*, available on: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168669.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*(4312), 75-78. doi:10.1126/science.198.4312.75

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57-111. doi:10.1017/s0140525x10000968

Minsky, M. (1974). A framework for representing knowledge. In J. Haugeland (Ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence* (pp. 95-128). Cambridge, MA: MIT Press.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science, 27*(12), 1550-1561. doi:10.1177/0956797616661199

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130487. doi:10.1098/rstb.2013.0487

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: Virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences, 18*(10), 512-519.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*(5), 1313-1326. Retrieved from <Go to ISI>://WOS:A1995TP40600019

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48-49.

Nash, J. (1953). Two-person cooperative games. *Econometrica, 21*(1), 128-140. doi:10.2307/1906951

Newton, E. L. (1990). *The Rocky Road from Actions to Intentions*. PhD dissertation: Stanford University.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford, UK: Oxford University Press.

Oaksford, M., & Chater, N. (1998). *Rationality in an Uncertain World*. Hove, England: Psychology Press.

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford, UK: Oxford University Press.

Olsson, E. (2021). Coherentist Theories of Epistemic Justification. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Vol. Spring 2021 Edition, pp. https://plato.stanford.edu/archives/spr2021/entries/justep-coherence/).

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., . . . Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology, 26*(10), 1334-1338. doi:10.1016/j.cub.2016.03.047

Paolini, S., Harwood, J., Hewstone, M., & Neumann, D. L. (2018). Seeking and avoiding intergroup contact: Future frontiers of research on building social integration. *Social and Personality Psychology Compass, 12*(12). doi:10.1111/spc3.12422

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica, 52*(4), 1029-1050. doi:10.2307/1911197

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169-190. Retrieved from <Go to ISI>://WOS:000224740800001

Pickering, M. J., & Garrod, S. (2021). *Understanding Dialogue: Language Use and Social Interaction*. Cambridge, UK: Cambridge University Press.

Pinel, E. C., Long, A. E., Landau, M. J., Alexander, K., & Pyszczynski, T. (2006). Seeing I to I: A pathway to interpersonal connectedness. *Journal of Personality and Social Psychology, 90*(2), 243-257. doi:10.1037/0022-3514.90.2.243

Povinelli, D. J., & Eddy, T. J. (1996). Chimpanzees: Joint visual attention. *Psychological Science, 7*(3), 129-135. doi:10.1111/j.1467-9280.1996.tb00345.x

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*(2), 243-256. doi:10.1037/0022-3514.64.2.243

Pylyshyn, Z. W. (Ed.) (1987). *The Robot's Dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Quine, W. V. O., & Ullian, J. S. (1978). *The Web of Belief*. New York: Random House.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin, 121*(1), 133-148. doi:10.1037/0033-2909.121.1.133

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory, 25*(1), 92-100. doi:10.1016/0022-0531(81)90018-1

Rossignac-Milon, M., Bolger, N., Zee, K. S., Boothby, E. J., & Higgins, E. T. (2021). Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology, 120*(4), 882-911. doi:10.1037/pspi0000266

Rousseau, D. M. (1989). Psychological and implied contracts in organizations. *Employee Responsibilities and Rights Journal, 2*(2), 121-139.

Rousseau, D. M. (1995). *Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements*. London: Sage Publications.

Rousseau, D. M., & Shperling, Z. (2003). Pieces of the action: Ownership and the changing employment relationship. *Academy of Management Review, 28*(4), 553-570. doi:10.5465/amr.2003.10899368

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Erlbaum.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schiffer, S. R. (1972). *Meaning*. Oxford: Oxford University Press.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*(4), 393-414. doi:10.1017/s0140525x12000660

Schmid, H. B. (2012). Shared intentionality and the origins of human communication. In A. Salice (Ed.), *Intentionality: Historical and Systematic Perspectives* (pp. 349-368). Munich: Philosophia Verlag.

Schmidt, M. F. H., & Tomasello, M. (2012). Young children enforce social norms. *Current Directions in Psychological Science, 21*(4), 232-236. doi:10.1177/0963721412448659

Scholl, A., & Sassenberg, K. (2014). Where could we stand if I had ... ? How social power impacts counterfactual thinking after failure. *Journal of Experimental Social Psychology, 53*, 51-61. doi:10.1016/j.jesp.2014.02.005

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive and reconstructive power of social norms. *Psychological Science, 18*(5), 429-434. doi:10.1111/j.1467-9280.2007.01917.x

Searle, J. R. (1990). Collective intentions and actions. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in Communication* (pp. 401-415). Cambridge, MA: MIT Press.

Searle, J. R. (1995). *The Construction of Social Reality*. New York: Free Press.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70-76. doi:10.1016/j.tics.2005.12.009

Sellars, W. (1968). *Science and Metaphysics: Variations on Kantian Themes*. London: Routledge and Kegan Paul.

Sennet, A. (2016). Ambiguity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Vol. available at: https://plato.stanford.edu/archives/spr2016/entries/ambiguity/).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55-89. doi:10.1016/j.cogpsych.2013.12.004

Sherif, M. (1936). *The Psychology of Social Norms*. New York: Harper.

Shteynberg, G., & Apfelbaum, E. P. (2013). The power of shared experience: Simultaneous observation with similar others facilitates social learning. *Social Psychological and Personality Science, 4*(6), 738-744. doi:10.1177/1948550613479807

Shteynberg, G., Hirsh, J. B., Apfelbaum, E. P., Larsen, J. T., Galinsky, A. D., & Roese, N. J. (2014). Feeling more together: Group attention intensifies emotion. *Emotion, 14*(6), 1102-1114. doi:10.1037/a0037697

Shteynberg, G., Hirsh, J. B., Bentley, R. A., & Garthoff, J. (2020). Shared worlds and shared minds: A theory of collective learning and a psychology of common knowledge. *Psychological Review, 127*(5), 918-931. doi:10.1037/rev0000200

Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews, 30*(6), 855-863. doi:10.1016/j.neubiorev.2006.06.011

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling Communication to Coordinate Perspectives in Cooperation. *arXiv preprint, arXiv:2106.02164*.

Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization, 25*(3), 309-327. doi:10.1016/0167-2681(94)90103-1

Sugden, R. (1989). Spontaneous order. *Journal of Economic Perspectives, 3*(4), 85-97. doi:10.1257/jep.3.4.85

Sugden, R. (2003). The logic of team reasoning. *Philosophical explorations, 6*(3), 165-181.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information, 13*(2), 65–93. Retrieved from <Go to ISI>://A1974T197600004

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201-212. doi:10.1016/j.tics.2017.12.005

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285. doi:10.1126/science.1192788

Thomas, K. A., DeScioli, P., Haque, O. S., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology, 107*(4), 657-676. doi:10.1037/a0037037

Thomas, R., Sargent, L. D., & Hardy, C. (2011). Managing organizational change: Negotiating meaning and power-resistance relations. *Organization Science, 22*(1), 22-41. doi:10.1287/orsc.1090.0520

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. In S. T. Fiske (Ed.), *Annual Review of Psychology* (Vol. 66, pp. 519-545).

Tomasello, M. (2003). *Constructing a Language*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences, 43*(e56), 1-58. doi:https://doi.org/10.1017/S0140525X19001742

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*(1), 121-125. doi:10.1111/j.1467-7687.2007.00573.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691. doi:10.1017/s0140525x05000129

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind & Language, 18*(2), 121-147. doi:10.1111/1468-0017.00217

Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. In H. Høgh-Olesen (Ed.), *Human Morality and Sociality: Evolutionary and Comparative Perspectives* (pp. 91-234). London, UK: Palgrave-Macmillan.

Tuomela, R. (2005). We-intentions revisited. *Philosophical Studies, 125*(3), 327-369. doi:10.1007/s11098-005-7781-1

Tuomela, R., & Miller, K. (1988). We-intentions. *Philosophical Studies, 53*(3), 367-389. doi:10.1007/bf00353512

Turner, K., & Horn, L. (Eds.). (2018). *Pragmatics, truth and underspecification: Towards an atlas of meaning*. Leiden, NL: Brill.

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences, 3*(3), 373-381. doi:10.1017/s0140525x0000546x

Vanderschraaf, P. (1998). Knowledge, equilibrium and convention. *Erkenntnis, 49*(3), 337-369.

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, 23*(8-9), 998-1003. doi:10.1016/j.neunet.2010.06.002

Wang, R., Wu, S., Evans, J., Tenenbaum, J., Parkes, D., & Kleiman-Weiner, M. (2021). Too Many Cooks: Coordinating multi-agent collaboration through inverse planning. In S. Muggleton & N. Chater (Eds.), *Human-like Machine Intelligence* (pp. 152-170). Oxford, UK: Clarendon Press.

Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy, 11*(3), 271-294. doi:10.1111/j.1532-7078.2007.tb00227.x

Weiner, B. (2018). The legacy of an attribution approach to motivation and emotion: A no-crisis zone. *Motivation Science, 4*(1), 4-14.

Wellman, H. M. (1992). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Weymark, J. A. (2016). Social welfare functions. In *The Oxford Handbook of Well-Being and Public Policy* (pp. 1667). Oxford: Oxford University Press.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1528), 2417-2428. doi:10.1098/rstb.2009.0069

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128. doi:10.1016/0010-0277(83)90004-5

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308. doi:10.1016/j.tics.2006.05.002

Ziegler, J., Bencini, G., Goldberg, A., & Snedeker, J. (2019). How abstract is syntax? Evidence from structural priming. *Cognition, 193*. doi:10.1016/j.cognition.2019.104045

Area of psychology	Social perception: Understanding the behavior of others	Social influence: Shaping the thoughts and behavior of others	Social transmission: The propagation of thoughts and behavior	Social interaction: Two-way interplay of agent's behavior
Social psychology	Attribution theory (Weiner, 2018) Interpreting emotions from faces (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015) Implicit processes in evaluating others (Greenwald & Lai, 2020)	Obedience (Grzyb, Dolinski, Trojanowski, & Bar-Tal, 2018) Conformity (Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007) Mechanisms of persuasion (Cialdini, 2001) Bystander effects (Fischer et al., 2011)	Automatic imitation (Cracco et al., 2018) Emotional contagion (Hatfield, Cacioppo, & Rapson, 1993) Transmission of beliefs and stories (Heath et al., 2001)	Role-play experiments (Haney, Banks, & Zimbardo, 1973) Impact of shared attention on cognition, emotion and affiliation (Haj-Mohamadi et al., 2018)
Social and cognitive development	"Theory of mind" tasks (Wimmer & Perner, 1983) Inferring goals (Gergely, Nadasdy, Csibra, & Biro, 1995; Hamlin, Wynn, & Bloom, 2007)	Socialization of moral norms (Grusec & Goodnow, 1994) Children enforcing social norms on others (Schmidt & Tomasello, 2012)	Imitation of facial expressions in neonates (Meltzoff & Moore, 1977; Oostenbroek et al., 2016) Imitating aggressive behavior (Bandura, Ross, & Ross, 1961) Rational imitation in infants (Gergely, Bekkering, & Kiraly, 2002)	Early shared intentionality (Tomasello & Carpenter, 2007) Helping behavior in infants (Warneken & Tomasello, 2007)

Comparative cognition	Chimps infer food location from gaze (Hare, Call, & Tomasello, 2001) Scrub-jays food-caching depends on what others have observed (Emery & Clayton, 2001)	Machiavellian intelligence: chimps apparently attempt to mislead others on location of food (Byrne & Whiten, 1989) Animal signaling as manipulation (Dawkins & Krebs, 1978) Collection action in animal groups (Couzin, Krause, Franks, & Levin, 2005)	Imitation and emulation (Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009) Cultural learning (Heyes & Galef Jr, 1996) Possible 'teaching' between animals (Hoppitt et al., 2008)	Joint attention (Povinelli & Eddy, 1996) Joint action (Bullinger, Wyman, Melis, & Tomasello, 2011) Cooperative hunting (Boesch, 1994)

Table 1. Social perception, influence, transmission, and interaction in studies in social, developmental, and comparative psychology.

While interaction seems central to social behavior, many fields of research focus on one-directional relationships between people: how people interpret, influence, or transmit information or behavior to each other. These phenomena can be studied without facing the problem of mutual interdependence that generates the paradox of social interaction: that in social interaction each person is trying to second-guess the thought and behavior of the other.

Type of theory	Outline	Challenge from mutually interdependent interaction	Illustrative references
Theory-theory	Each person formulates a 'theory' of the other's beliefs and desires	Circularity: A's theory of B includes B's theory of A	Gopnik and Wellman (1992); Wellman (1992)
Simulation theories	Each person uses their own mind to simulate another's	Circularity: A simulates B who is simulating A	Goldman (2006); Gordon (1986)
Prediction-based theories	Each person is a "prediction machine" who best-responds to what she predicts the other will do	Circularity: A's choice depends on its prediction of B's choice; B's choice depends on its prediction of A's choice	Tamir and Thornton (2018)
Truncated recursive theories	Cognitive hierarchy theory; k-level reasoning; higher-order intentionality	Depends on heuristics concerning the "0th" recursive level	Camerer et al. (2004); Stahl and Wilson (1994)
Bayesian models of mutual prediction	Softmax choice rule and iterate to find "fixed points"	Problem of choosing between multiple equilibria	Shafto et al. (2014)
Rational speech act theory	Iterate to find "fixed points"	Problem of choosing between multiple equilibria	Frank and Goodman (2012)
Nash equilibrium	Each person best-responds to the other's strategy	Problem of choosing between multiple equilibria	Nash (1950)
Rationalizability	Each person optimizes in light of their own beliefs. Solving for "fixed points"	Problem of choosing between multiple choice vectors	Bernheim (1984); Pearce (1984)

Table 2. Individualistic approaches to the problem of reasoning about social interaction in psychology, philosophy, and economics.